

University of Kentucky

UKnowledge

---

Theses and Dissertations--Accountancy

Accountancy

---


2021

## Earnings Conference Calls and Lazy Prices

Chuancai Zhang

University of Kentucky, zcc198621@gmail.com

Author ORCID Identifier:

 <https://orcid.org/0000-0003-3404-168X>

Digital Object Identifier: <https://doi.org/10.13023/etd.2021.118>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

Zhang, Chuancai, "Earnings Conference Calls and Lazy Prices" (2021). *Theses and Dissertations--Accountancy*. 15.

[https://uknowledge.uky.edu/accountancy\\_etds/15](https://uknowledge.uky.edu/accountancy_etds/15)

This Doctoral Dissertation is brought to you for free and open access by the Accountancy at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Accountancy by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Chuancai Zhang, Student

Dr. Hong Xie, Major Professor

Dr. Brian Bratten, Director of Graduate Studies

EARNINGS CONFERENCE CALLS AND LAZY PRICES

---

DISSERTATION

---

A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy in the  
College of Business and Economics  
at the University of Kentucky

By  
Chuancai Zhang  
Lexington, Kentucky  
Director: Dr. Hong Xie, Professor of Accountancy  
Lexington, Kentucky  
2021

Copyright © Chuancai Zhang 2021  
<https://orcid.org/0000-0003-3404-168X>

## ABSTRACT OF DISSERTATION

### EARNINGS CONFERENCE CALLS AND LAZY PRICES

Changes to the language and construction of financial reports are indicative of firms' future returns and operations. Investors, however, are often inattentive to these changes; consequently, price reactions to these changes are delayed—resulting in “lazy” prices. This study explores two possible channels through which earnings conference calls may mitigate lazy prices: (1) the topic overlap channel and (2) the comparison language channel. Specifically, I examine whether the topic overlap between conference call transcripts and 10K/10Q filings or the comparison language used on earnings conference call transcripts helps investors understand the nature of the overlapped topics and triggers investors' attention to firms' financial reports and changes therein, and thus attenuates the predictive ability of changes in textual narratives for future returns. The main results support both channels. Further analysis shows that the comparison language in both the Q&A session and the Presentation session of earnings conference calls helps mitigate lazy prices. This study contributes to both the earnings conference call literature and the capital market literature by providing new insights into the role of earnings conference calls in the capital markets and contributes to textual analysis literature by providing best practices of applying topic modeling methods to accounting research.

**KEYWORDS:** Financial Reports, Document Similarity, Topic Overlap, Comparison Language, Earnings Conference Calls, Lazy Prices

---

Chuancai Zhang  
*(Name of Student)*

---

05/04/2021

Date

EARNINGS CONFERENCE CALLS AND LAZY PRICES

By  
Chuancai Zhang

Dr. Hong Xie  
\_\_\_\_\_  
Director of Dissertation

Dr. Brian Bratten  
\_\_\_\_\_  
Director of Graduate Studies

05/04/2021  
\_\_\_\_\_  
Date

## ACKNOWLEDGMENTS

I would like to thank my dissertation committee members, Hong Xie, Dan Stone, Russell Jame, and Xin Ma, for their guidance and support. I would also like to thank Vikram Gazula and Satrio Husodo for their amazing technology support, and the faculty and doctoral students at Von Allmen School of Accountancy for their insightful feedbacks. I am so thankful to everyone who has supported me along this journey. I gratefully acknowledge the financial support from the University of Kentucky, the Gatton College of Business, and the Von Allmen School of Accountancy. All errors are my own.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS .....	iii
LIST OF TABLES .....	vi
LIST OF FIGURES .....	vii
CHAPTER 1. INTRODUCTION .....	1
CHAPTER 2. LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT.....	8
2.1 Literature on the Usefulness of Financial Reports.....	8
2.2 Literature on the Usefulness of Earnings Conference Calls .....	9
2.3 Hypotheses Development .....	11
2.3.1 Topic Overlap and Lazy Prices .....	13
2.3.2 Comparison Language and Lazy Prices.....	14
CHAPTER 3. RESEARCH DESIGN.....	17
3.1 Sample and Data .....	17
3.2 Variable Measurement .....	22
3.2.1 Document Similarity .....	22
3.2.2 Topic Overlap and Comparison Language .....	23
3.3 Method .....	25
3.3.1 Confirmation of the Existence of Lazy Prices .....	25
3.3.2 Effect of Topic Overlap and Comparison Language on Lazy Prices.....	26
3.4 Validation of the Topic Modeling Approaches.....	26
CHAPTER 4. EMPIRICAL RESULTS.....	28
4.1 Descriptive Statistics.....	28
4.2 Correlation Analysis .....	30
4.3 Confirmation of the Existence of Lazy Prices .....	32
4.4 The Effect of Topic Overlap on Lazy Prices (H1).....	35
4.5 The Effect of Comparison Language on Lazy Prices (H2).....	37
CHAPTER 5. ADDITIONAL ANALYSIS.....	40
5.1 Comparison Language in the Presentation Session and Q&A Session .....	40
5.2 Document Similarity based on Word Stem.....	42

CHAPTER 6. CONCLUSION.....	45
APPENDICES .....	47
APPENDIX 1. VARIABLE DEFINITIONS.....	47
APPENDIX 2. DOCUMENT SIMILARITY AND LATENT DIRICHLET ALLOCATION.....	49
APPENDIX 3. INDUSTRY TOPICS SUMMARY .....	69
REFERENCES .....	94
VITA.....	98



## LIST OF TABLES

Table 3.1 Sample Selection Process and Sample Distribution .....	20
Table 4.1 Descriptive Statistics.....	30
Table 4.2 Correlation Matrix for the Main Variables .....	31
Table 4.3 Test of Lazy Prices Based on Cohen et al. (2020) Table IV (1995-2014) .....	33
Table 4.4 Extension of Cohen et al. (2020) Table IV (1995-2018).....	34
Table 4.5 The Effect of Topic Overlap on Lazy Prices .....	36
Table 4.6 The Effect of Comparison Language on Lazy Prices .....	38
Table 5.1 The Effect of Comparison Language on Lazy Prices (Q&A VS. Presentation).....	40
Table 5.2 The Effect of Document Similarity on a Firm's Future Return (Stem) .....	44

## LIST OF FIGURES

Figure 3.1 The Number of Firms in Each Calendar Quarter (1995-2018) .....	17
Figure 3.2 Earnings Announcement Date and Earnings Conference Call Date Distribution around 10K/10Q Filing Date .....	19

## CHAPTER 1. INTRODUCTION

Cohen, Malloy, and Nguyen (2020) find that changes in textual narratives of financial reports (year-to-year changes for annual reports or 10Ks and seasonal changes for quarterly reports or 10Qs) contain rich information about future returns and operations. Investors, however, often ignore (or are inattentive to) textual changes when financial reports are released, as indicated by little stock price reactions around the financial report release date. Investors respond to these textual changes only gradually over time, i.e., stock prices impound the information embedded in textual changes gradually. Cohen et al. (2020) term delayed price responses to textual changes as “lazy prices”.

In this study, I examine whether earnings conference calls can mitigate lazy prices. I examine the role of earnings conference calls because some conference calls may directly discuss the important information in the financial reports, or indirectly alert investors to the financial reports and changes therein. Specifically, I explore two possible channels through which earnings conference calls may mitigate lazy prices. The first channel is that the discussion of overlapped topics between earnings conference calls and 10K/10Q filings helps investors understand the overlapped topics, which are otherwise ignored (Cohen et al., 2020), and their changes compared to the prior year. Because prior literature shows that investors respond to the information released in conference calls (Lee, 2016; Mayew, Sethuraman, and Venkatachalam, 2020), the information in these overlapped topics are impounded into stock prices at conference calls before the release of the 10K/10Q filings. Consequently, the narrative changes related to these overlapped topics can no longer predict future return. I call this channel the “topic overlap channel.”

The second channel is that the comparison language (such as “compared to” and “previous year”) used in the earnings conference calls can attract investors’ attention to a firm’s previous year and current year’s financial reports. Therefore, the information in the narrative changes of the financial reports can be impounded into prices in a timely manner and the lazy price is mitigated. I call this channel the “comparison language channel.”

This study focuses on earnings conference calls for several reasons. First, earnings conference calls may be the closest event to 10K/10Q releases. Earnings conference calls are usually held on the same day as earnings announcements or one day later. Increasingly firms release financial reports concurrently with earnings announcements (Arif, Marshall, Schroeder and Yohn, 2019). Therefore, the earnings conference call date is very close to the financial report date. Second, firms believe that public earnings conference calls are the most important channel to convey company messages to investors, especially institutional investors, and provide opportunities for analysts to ask some questions which are directly from institutional investors during the conference calls (Brown, Call, Clement and Sharp, 2019). Both analysts and fund managers regard earnings conference calls as important for generating earnings forecasts or better understanding the firm (Brown, Call, Clement, and Sharp, 2015; Barker, Hendry, Roberts, and Sanderson, 2012). Kimbrough (2005) find the initiation of conference calls is associated with a significant reduction in analysts' and investors' underreaction to the future implications of currently announced earnings. Therefore, earnings conference calls may draw both institutional and retail investors’ attention. Third, earnings conference calls have two sessions, the presentation session and Q&A session. In the Presentation session, firms have opportunities to further disclose information to investors, and in the Q&A session, analysts can ask management

questions that may interest investors. Therefore, conference calls provide a setting to examine both the topics discussed and the language used by firm executives and analysts.

To explore the topic overlap channel, I create a corpus with 469,918 10K/10Q filings and 107,413 earnings conference call transcripts. I apply the Latent Dirichlet Allocation (LDA) topic modeling method to this corpus and extract topics for each document. Then I create two topic overlap measures based on the topics in each pair of 10K/10Q filings and the earnings conference call transcripts. The first topic overlap measure is the topic based Jaccard similarity (TopicOverlap) which captures the extent of overlap between the topics in the 10K/10Q filings and the earnings conference call transcripts. The second topic overlap measure is the ratio of the number of overlapped topics in the 10K/10Q filings and earnings conference call transcripts to the number of topics in the 10K/10Q filings (TopicCoverage). Based on the Fama-MacBeth cross-sectional regressions model used in Cohen et al. (2020), which regresses individual firm-level future stock returns on the document similarity measure, and control variables, I interact the two topic overlap measures with the document similarity measure separately to test this channel. A significant negative coefficient on the interaction indicates that the topic overlap between the 10K/10Q filings and earnings conference call transcripts mitigates the lazy prices. The results support this channel, specifically the more the topics overlap between the 10K/10Q filings and earnings conference call transcripts, the less “lazy” are the prices.

To explore the comparison language channel, I create a comparison language measure (CompWord) that captures the number of comparison words and phrases used on conference call transcripts. The comparison words and phrases include “previous”,

“compared”, “last year”, “prior year”, “previous year”, “compared to” and “compared with.” Specifically, I calculate the comparison language measure by using the log of the ratio of the number of comparison words and phrases in each earnings conference call transcript to the document length of the transcript multiplied by 10000. Like the test of the topic overlap channel, I interact the comparison language measure with the document similarity measure in the Fama-MacBeth cross-sectional regressions model used in Cohen et al. (2020). The results show that while the coefficient on the interaction between the comparison language measure and the cosine similarity measure is insignificant, the coefficient on the interaction between the comparison language measure and the Jaccard similarity measure is negative and marginally significant at 10%. When I narrow down the sample to include only observations that the day-difference between their earnings announcement date and 10K/10Q releasing date is within the range [-5,0]. The significance level of the coefficient on the interaction between the comparison language measure and the Jaccard similarity measure increases from 10% to 1% though the coefficient on the interaction between the comparison language measure and the cosine similarity measure is still insignificant. The result indicates that due to investors’ limited attention, the comparison language channel is more pronounced when the earning conference call date is closer to the 10K/10Q releasing date. I further examine whether the comparison language used in the Presentation session and the Question & Answer session has different effects on investors. I find that the comparison language in both the Q&A session and Presentation session can help mitigate the lazy prices. Overall, the results support the comparison language channel that the use of comparison language on earnings conference

call transcripts can attract investors' attention to firms' current year and prior year's financial reports and thus mitigates lazy prices.

This study contributes to the literature and practice in several important ways. First, this study contributes to the earnings conference call literature. Earnings conference calls have become a prevalent voluntary disclosure medium for U.S. firms (Skinner, 2003; Bushee, Matsumoto, and Miller, 2004). However, critics argue that earnings conference calls—even the Q&A portion—often involve more “theater” than prior literature documented (Brown et al., 2019). If earnings conference calls are only a “show” between analysts and management, why do firms still frequently host quarterly earnings conference calls? Evidence has shown some direct effects of earnings conference calls, such as providing more information about the firm to investors and analysts (e.g., Mayew et al., 2020). However, can earnings conference calls help investors understand the information in the textual narratives of financial reports and attract investors' attention to the textual narratives of financial reports? To my knowledge, this study is the first to explore the two possible channels (the topic overlap channel and the comparison language channel) through which earnings conference calls help attenuate lazy prices. The metrics that I examine (the topic overlap between earnings conference calls and financial reports, and the comparison language used on earnings conference call transcripts) have not been examined in the earnings conference call literature.

Second, this study explores channels may potentially mitigate the “lazy prices” anomaly documented by Cohen et al. (2020)—simple changes in textual narratives of financial reports can predict future returns but there is no announcement effect of financial reports when released due to investors' inattention. Regulations have required firms to

disclose more textual (non-accounting) information, such as internal control, MD&A, and risk factors, in financial reports to help investors make better decisions. However, if textual information in financial reports is neither relevant nor informative, why burden firms with preparing this information? Literature shows that textual information in financial reports has value relevance, but investors initially miss the information when financial reports are first released and only uncover the information gradually over time (Cohen et al., 2020). My findings show the “topic overlap channel” of earnings conference calls help investors understand the text information in the financial reports and the “comparison language channel” of earnings conference calls alert investors’ attention to the financial reports and thus mitigate lazy prices.

Finally, this study contributes to the textual analysis literature in accounting by providing best practices of applying topic modeling methods to accounting research. LDA is a popular method in topic modeling and a few accounting studies have applied LDA to extract topics (i.e., Brown, Crowley, and Elliott, 2020; Calomiris and Mamaysky, 2019; Huang, Leavy, Zang, and Rong, 2018; Gomez, Heflin, Lee, and Wang, 2018; Dyer, Lang, and Stice-Lawrence, 2017). However, LDA has some substantive disadvantages, one of which is that it is non-deterministic.<sup>1</sup> Therefore, without careful design, the results from LDA may be misleading. But extant studies rarely discuss the disadvantages. This study compares the topic coherence and topic stability of different LDA models in each industry to determine the preferred model in this study. A model that can generate topics with high topic coherence and topic stability is preferred. This study also compares the performance

---

<sup>1</sup> LDA uses randomness in training the model and each time it generates different topics even for the same corpus. Therefore, it is important to carefully choose a model that is much more stable in terms of the resulting topics.



of the Gensim's LdaMulticore model with that of the Tomotopy's LDA model by applying them to SEC 10K/10Q filings and earnings conference call. Finally, this study gives several suggestions to scholars who are interested in applying topic modeling methods to accounting research.

This study proceeds as follows. Section 2 discusses the related literature and develops the hypothesis. Section 3 describes the research design. Section 4 presents the main empirical results. Section 5 shows the additional analysis results. Section 6 concludes this study.

## CHAPTER 2. LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

### 2.1 Literature on the Usefulness of Financial Reports

Financial reports are important channels for investors to get firm information and provide benchmarks and fundamental signals that may indicate future performance. Financial reports contain both numerical accounting information (e.g., sales revenue and earnings) and textual narratives (e.g., footnotes, MD&As, and risk factor disclosures). On the one hand, evidence shows that the relevance of accounting information has deteriorated markedly (Brown, Lo, and Lys, 1999). Reported earnings no longer provide a reliable basis to predict firm's future performance. For example, Lev and Gu (2016) find that although corporate information has an increasing impact on investors' decisions, only a small amount of that information is contributed by companies' quarterly and annual reports.

On the other hand, though the fast-growing textual analysis literature show that textual narratives in financial reports are also informative/useful to investors (Li, 2010; Lee, 2012; Feldman, Govindaraj, Livnat, and Segal, 2010; Ertugrul, Lei, Qiu, and Wan, 2017), the usefulness of textual narratives has also decreased over time. For example, Brown and Tucker (2011) find that the MD&A modification is informative to the capital market as the stock price responses to 10-K filings is positively associated with the MD&A modification. However, the usefulness of the MD&A modification has declined as the price reaction to the MD&A modification scores has weakened in the past decade.

However, Cohen et al. (2020) argue that changes in textual narratives of financial reports contain rich information that predict future performance or returns, but investors pay little attention to financial reports upon their release. Accounting reports still have information, but investors only slowly find the value of the text information in accounting

reports. Drake, Roulstone, and Thornock (2015) find that investors continue to acquire historical accounting reports long after their release, an activity that occurs frequently.

If investors' inattention leads to no or little reaction to 10Q/10K releases, what factors can draw investors' attention to financial reports? Drake, Roulstone, and Thornock (2016) studied four factors that drive investors to search for historical accounting reports. They find that requests for historic reports are positively associated with financial reporting complexity, accounting discretion, negative earnings shocks, and shocks to firm value (particularly negative shocks). But they show that investors seek out these historical accounting reports because they contain qualitative and quantitative information that helps contextualize current-period information and is useful for current-period decision making. This shows the confirmative role of financial reports.

Considering the predictive value of financial reports, what factors may help investors understand and impound the information in the textual narratives of financial reports into price in a timely manner? Can earnings conference calls play a role?

## 2.2 Literature on the Usefulness of Earnings Conference Calls

Earnings conference calls have become a prevalent voluntary disclosure medium for U.S. firms as most public firms regularly host quarterly earnings conference calls (Skinner, 2003; Bushee et al., 2004). Early research focuses on the earnings conference call itself and establishes that conference calls are informative to market participants in that they trigger heightened trading and stock price responses and help analysts form more accurate earnings expectations (Bowen et al. 2002).<sup>2</sup> Brown, Hillegeist and Lo (2004) find

---

<sup>2</sup> Bowen, Davis, and Matsumoto (2002) find that conference calls increase analysts' ability to forecast earnings accurately, suggesting that these calls increase the total information available about a firm and conference calls help 'level the playing field' across analysts.

that earnings conference calls can reduce information asymmetry and the cost of capital. Kimbrough (2005) find the initiation of conference calls is associated with a significant reduction in analysts' and investors' underreaction to the future implications of currently announced earnings. Both analysts and fund managers regard earnings conference calls as important for generating earnings forecasts or better understanding the firm (Brown et al., 2015; Barker et al., 2012). Analysts and investors benefit from asking follow-up questions, requesting more details, and perhaps questioning management's interpretation of events (Matsumoto, Pronk, and Roelofsen, 2011).

Recently, taking advantage of the development of machine learning and natural language processing (NLP), textual analysis of earnings conference calls transcripts show that both the content and the interactive nature are informative. Extant studies mainly focus on the market consequences (such as market reaction, abnormal return, stock return volatility) of specific linguistic characteristics of top executives, such as readability or complexity (Brochet, Loumiot, and Serafeim, 2015; Burgoon et al., 2015; Bushee, Gow, and Taylor, 2018), tone and tone dispersion (Davis, Ge, Matsumoto, and Zhang, 2015; Allee and Deangelis, 2015), lack of spontaneity (Lee, 2016), linguistic opacity (Brochet, Naranjo, and Gwen, 2016), and language vagueness (Dzielinski, Wagner, and Zeckhauser, 2016). With more high frequent stock trading data available to researchers, the intra-day capital market reactions to conference call characteristics become more precise and convincing. Mayew et al. (2020) uses intra-day absolute stock price reactions around specific analyst-manager dialogues to measure informativeness. They find that manager dialogues with disfavored analysts are more informative and stock prices directionally respond to both the analyst's linguistic tone and the manager's voice pitch.

Despite the informativeness of conference calls, prior literature also criticizes that the conference calls are heavily manipulated by firms. Brown et al. (2019) conducts a survey of 610 investor relations officers (IROs) and 14 follow-up interviews. They find that the IROs believe that public earnings conference calls are the single most important tool for conveying the company message to institutional investors. IROs help managers carefully manage every aspect of these calls, including “developing a script, preparing a list of possible questions and answers, developing a strategy for handling unanticipated questions, and rehearsing the call.” And they also control who can ask questions during the conference call.

In summary, despite the finding that companies may manipulate earnings conference calls, both the content and the interactive nature of earnings conference calls are informative to investors. This study examines whether earnings conference calls attenuate lazy prices documented in Cohen et al. (2020) if conference calls cover some topics that are in financial reports or contain comparison languages.

### 2.3 Hypotheses Development

Cohen et al (2020) find changes in textual narratives of the 10K/10Qs predict future performance or returns, but investors are inattentive to the text changes of the financial reports when released. Investors uncover the implications of changes in textual narratives only gradually over time, but eventually the news is fully impounded into stock prices and reflected in firm operations. The question here is why investors are inattentive to financial reports upon their release?

One explanation is that investors only find earnings information but not text information useful. As investors already get earnings information at earnings

announcements, they do not pay attention to financial reports upon their release because they believe financial reports contain the same earnings information as earnings announcements. If so, investors will not use financial reports later. However, Drake et al. (2015) find that investors continue to acquire historical accounting reports long after their releases, and this activity occurs frequently. Therefore, this explanation cannot answer the question that why investors are inattentive to financial reports upon their release.

The second explanation is that investors believe financial reports are useful, but only find the confirmative value of financial reports. Drake et al. (2016) studied four factors that drive investors to search for historical accounting reports. They find that requests for historic reports are positively associated with financial reporting complexity, accounting discretion, negative earnings shocks, and shocks to firm value (particularly negative shocks). These results support the idea that investors seek out these historical accounting reports because they contain qualitative and quantitative information that helps contextualize current-period information and is useful for current-period decision making. Therefore, investors only use financial reports to confirm the information they already got.

Consistent with the second explanation, if investors seek the predictive value of financial reports, they will identify changes in textual narratives of financial reports and impound the information embedded in changes in textual narratives to prices in a timely manner, i.e., when financial reports are released. However, if investors only seek confirmative value of financial reports, then they will explore financial reports but not in a timely manner.

However, evidence shows that investors also identify predictive roles of the financial reports, but that the predictive value of financial reports is low. Feldman et al.

(2010) find that a positive tone in the MD&A section is associated with modestly higher contemporaneous and future returns and that an increasingly negative tone is associated with lower contemporaneous returns. Brown and Tucker (2011) find that the magnitude of stock price responses to 10-K filings is positively associated with the MD&A modification score, but the price reaction to MD&A modification scores has weakened over time.

The lazy prices anomaly identified by Cohen et al. (2020) is likely due not only to investors' inattention to financial report releases, but also to investors' perception of the role of financial reports. This study explore two possible channels through which earnings conference calls may mitigate lazy prices.

### 2.3.1 Topic Overlap and Lazy Prices

The first channel is that the discussion of overlapped topics between earnings conference calls and 10K/10Q filings prompts investors to respond to these overlapped topics and their changes with respect to the prior year during conference calls. As the information in these overlapped topics are already impounded into prices during conference calls before the release of the 10K/10Q filings, the narrative changes related to these overlapped topics can no longer predict future returns. That is, conference calls mitigate lazy prices through the topic overlap channel.

More specifically, during conference calls, managers may choose to focus on certain topics and analysts may ask questions about these topics. The topics discussed during conference calls draw investors' attention, and are incorporated into stock price immediately, which is the inter-day market reaction to earnings conference calls (Mayew et al., 2020). If this is true, there should be market reaction around the talk about the topics during a conference call. Evidence supports this argument. For example, Gomez et al.

(2018) categorize conference call sentences into 29 topics and find that many topics such as regulations, risk, and competition cause large stock price movements with little drift in the immediate period following each sentence.

Therefore, I assume that (1) the topics discussed during conference calls can help investors understand the nature of the topics including the topics themselves and their changes compared to the prior period and (2) the overlapped topics between conference calls and 10K/10Qs are impound into prices timely and quickly during conference calls. Then the “hidden” information in the financial reports and changes therein, which are otherwise ignored by investors, can be “discovered” in the earnings conference calls and be impound into prices immediately.

If the “hidden” information in the financial reports and changes therein is “discovered” in the earnings conference calls and impound into prices immediately, then there will be no lazy prices. Therefore, more topic overlap between earnings conference calls and 10K/10Qs is negatively related to lazy prices documented in Cohen et al. (2020). I formulate my first hypothesis below.

*H1: Topic overlap between earnings conference call transcripts and 10K/10Q filings mitigates lazy prices.*

### 2.3.2 Comparison Language and Lazy Prices

The second channel is that the comparison language used in the earnings conference calls can attract investors’ attention to a firm’s previous year and current year’s financial reports. Therefore, the information in the narrative changes of the financial reports can be impounded into price in a timely manner and the lazy price is mitigated if conference calls contain comparison languages. I define comparison language as sentences or paragraphs



that contains the comparative words or phrases such as “previous”, “compared”, “last year”, “prior year”, “previous year”, “compared to” and “compared with.” I term this channel the comparison language channel.

Even if topic overlap potentially mitigates lazy prices, not all topics in the 10K/10Qs are discussed during the earnings conference calls. Therefore, there are still undiscovered “hidden” information in the textual narrative changes of 10K/10Q filings. Limited attention theory shows that attention may become a scarce cognitive resource when individuals face a rich supply of information (Falkinger, 2008). The accounting and finance literatures find that not only investors but also financial analysts are subject to limited attention. For example, Louis and Sun (2010) find that investors are less attentive to Friday announcements and inattention affects investors' information processing even in merger announcements. Driskill, Kirk, and Tucker (2020) find that even financial analysts are subject to limited attention when they face with concurrent earnings announcements. Besides, Arif et al. (2019) document that firms are increasingly disclosing earnings announcements (EA) concurrently with the 10-K filing instead of first issuing a ‘stand-alone’ EA over time and they find a muted market reaction to concurrent EA/10-Ks relative to stand-alone EAs. Therefore, investors even have less attention to explore the “hidden” information in the textual narrative changes of 10K/10Q filings when they face the concurrent disclosure.

However, Cohen et al. (2020) find that the return predictability of changes in textual narratives of financial reports only exists in firms that do not make explicit textual comparisons to prior accounting reports, which means that comparison statements or languages can draw investors’ attention to a prior year’s financial reports and facilitate

investors' information processing.<sup>3</sup> Similarly, I predict that comparison statements or languages used during earnings conference calls can also draw investors' attention to current and the prior year's accounting reports and thus mitigate lazy prices. First, as earnings announcement date and 10K/10Q filing date become closer, earnings conference call date is also closer to the 10K/10Q filing date. Therefore, the effect of earnings conference calls may become more pronounced. Second, as earnings announcement date and 10K/10Q filing date become closer, both analysts and investors do not have enough time to explore the information in the 10K/10Q filings. Therefore, on the one hand, analysts may do not have more valuable questions about the information in the 10K/10Q filings to ask during the conference calls. On the other hand, even the information in the 10K/10Q filings is discussed in the earnings conference call, it may not be thoroughly discussed, which may require investors to further study the information. Third, comparison statements or languages such as "previous", "compared", "last year", "prior year", "previous year", "compared to" and "compared with" are more likely to arouse investors' interest and attract them to assign their limited attention to firm's current and the prior year's accounting reports and thus mitigate lazy prices. My second hypothesis is below.

*H2: The use of comparison language on earnings conference call transcripts mitigates lazy prices.*

---

<sup>3</sup> While financial reports present current period's accounting numbers (e.g., the balance sheet) along with prior periods' numbers, prior periods' textual narratives are not presented along with current period's textual narratives.

## CHAPTER 3. RESEARCH DESIGN

### 3.1 Sample and Data

I constructed the sample of U.S. publicly listed firms from a variety of data sources. I obtained the 10K/10Q filings from the “Stage One 10-X Parse Data” provided by Dr. Bill McDonald.<sup>4</sup> Then I further cleaned the 10K/10Q filing documents and calculated the year over year Cosine Similarity and Jaccard Similarity score for each pair of 10K/10Q filings from 1995 to 2018. I obtained 766,250 10K/10Q filings with similarity scores. Figure 3.1 shows the number of firms in each calendar quarter (1995-2018).

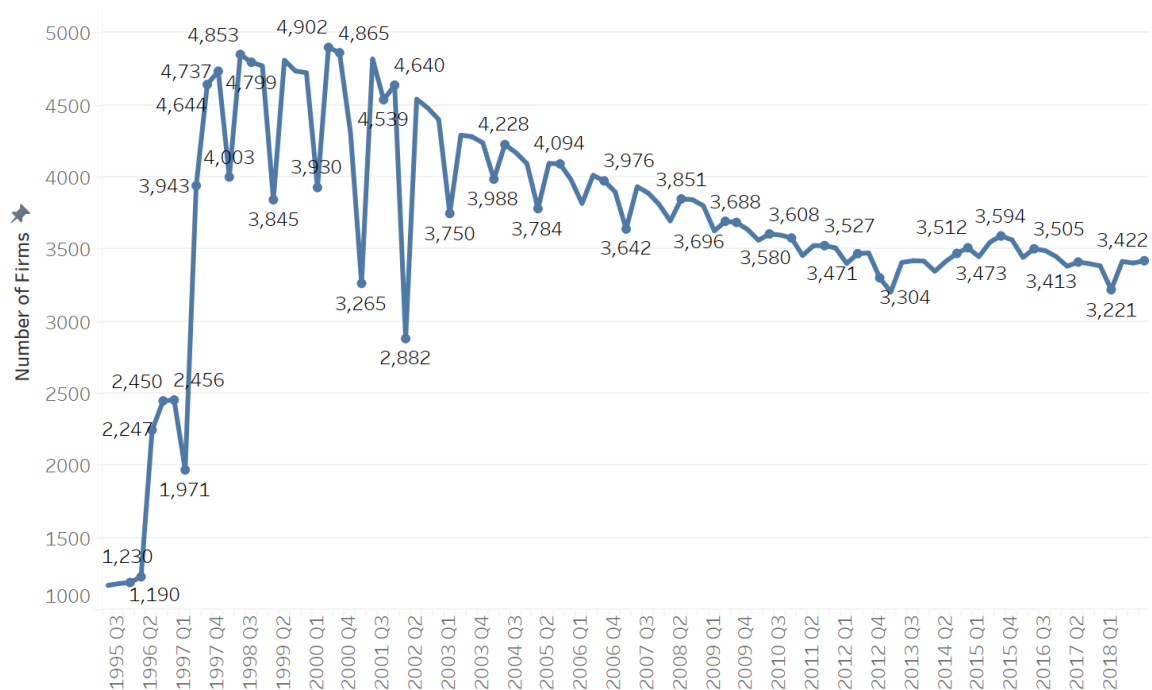


Figure 3.1 The Number of Firms in Each Calendar Quarter (1995-2018)

<sup>4</sup> This dataset is available from the website: <https://sraf.nd.edu/data/stage-one-10-x-parse-data/>. This dataset covers all the 10-X filings on the Security and Exchange Commission’s (SEC) EDGAR website from 1994 to 2018. The extraneous materials are removed from each filing document. Detailed information is available on the website. I choose this dataset instead of using the 10-X filings directly downloaded from SEC EDGAR website because this dataset is partially cleaned and available to all scholars. Using this dataset as a starting point, other scholars may find it easier to replicate my research.

I collected the earnings conference call transcripts from two data sources. I downloaded and cleaned 119,743 earnings conference call transcripts from Capital IQ from 2005-2018.<sup>5</sup> I collected and cleaned 89,988 firm quarter earnings conference call transcripts from SeekingAlpha.com between January 2005 and June 2017.<sup>6</sup> Though the two data sources cover different firms, some firms are overlapped in the two data sources. I utilized the Capital IQ as the main source and added 32,444 transcripts from SeekingAlpha.com that are not in the Capital IQ to the Capital IQ dataset. The final earnings conference call transcripts dataset contains 152,187 transcripts.

I obtained the financial data and stock return data used to calculate the main dependent variable and control variables from COMPUSTAT and CRSP.

I constructed the sample through the following process. Starting from the 766,250 10K/10Q filings (29,575 unique firms) with similarity scores, I matched the COMPUSTAT quarterly data using the CIK and filing date. I obtained 533,624 matched observations (16,624 unique firms). And then I matched the data to CRSP return data, removed observations with missing values on the main variables, and removed observations that the day-difference between earnings announcement date and 10K/10Q releasing date is not within the range [-60, 5]. I obtained a basic sample of 358,132 observations (11,558 unique firms) during 1995-2018. The sample from 1994-2014 is used to replicate the Table IV of Cohen et al. (2020). To test my main hypothesis, I further matched the data to earnings conference call transcripts data. I obtained a final sample with 199,911 observations (7,313

---

<sup>5</sup> The sample period starts from 2005 because of the availability of earnings conference call transcripts data in Capital IQ.

<sup>6</sup> Earnings conference call transcripts are available in Seeking Alpha website (<https://seekingalpha.com/earnings/earnings-call-transcripts>) and the starting year is 2005 for most of the firms. However, starting from 2018, Seeking Alpha updated their terms of use and stopped automatically downloading transcripts from the website.

unique firms) from 2005 to 2018. Figure 3.2 shows the earnings announcement date distribution and earnings conference call date distribution around 10K/10Q filing date.

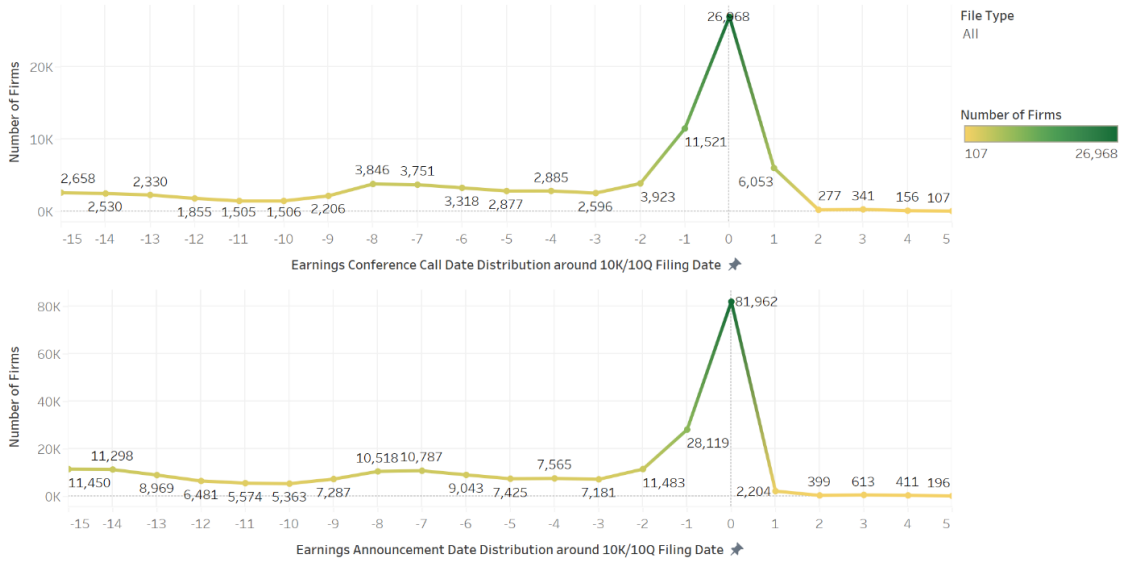


Figure 3.2 Earnings Announcement Date and Earnings Conference Call Date Distribution around 10K/10Q Filing Date

The sample selection process and the sample distribution based on calendar year and month is in Table 3.1.

Table 3.1 Sample Selection Process and Sample Distribution

Panel A: Summary of Sample Selection Process		
	# of observations	# of firms
Total observations with similarity score for pairs of 10K/10Q (1995-2018):	766,250	29,575
Less: Observations not matched with Compustat quarterly data (232,626/728,254=0.304)	(232,626)	
Total observations matched with Compustat quarterly data:	533,624	16,624
Less: Observations not matched with CRSP Return data	(145,916)	
Less: Observations with missing values on the main variables	(18,463)	
Less: Observations that the day-difference between earnings announcement date and 10K/10Q releasing date is not in the range (-60,5)	(11,113)	
Basic Sample (1995-2018):	358,132	11,558
Less: Observations not matched with earnings conference call data (2005-2018)	(157,455)	
Less: Observations that the day-difference between earnings conference call date and 10K/10Q releasing date is not in the range (-60,5)	(766)	
Final Sample for earnings conference call hypothesis tests (2005-2018):	199,911	7,313

Table 3.1 (Continued)

Panel B: Sample Distribution of the Basic Sample Based on Calendar Year and Month (1995-2018)													
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
1995	0	3	59	119	955	143	131	898	173	135	918	155	3,689
1996	123	378	753	362	1,783	258	328	1,805	382	335	1,824	352	8,683
1997	253	632	1,145	384	3,165	532	493	3,573	661	570	3,607	652	15,667
1998	359	1,140	2,594	601	3,944	609	457	3,771	652	417	3,841	603	18,988
1999	312	1,101	2,529	605	3,923	564	411	3,809	597	399	3,838	579	18,667
2000	319	1,103	2,601	551	4,076	551	390	3,920	616	426	3,910	4	18,467
2001	308	1,025	2,014	1,148	4,067	495	387	3,848	340	448	3,765	516	18,361
2002	273	965	1,719	1,053	3,772	444	381	3,651	537	409	3,592	486	17,282
2003	256	945	2,615	512	3,561	454	439	3,390	521	496	3,341	468	16,998
2004	261	953	2,846	607	3,365	465	474	3,245	514	424	3,262	470	16,886
2005	232	895	2,719	596	3,293	444	452	3,227	481	458	3,140	446	16,383
2006	232	971	2,679	474	3,306	416	410	3,218	407	454	3,089	414	16,070
2007	213	1,335	2,147	608	3,234	390	465	3,103	371	519	3,014	323	15,722
2008	187	1,843	1,705	568	3,069	367	538	2,980	371	650	2,829	360	15,467
2009	193	1,488	1,982	558	2,930	338	641	2,746	344	644	2,694	332	14,890
2010	189	1,534	1,873	674	2,706	317	602	2,708	328	577	2,719	315	14,542
2011	184	1,588	1,718	592	2,738	283	475	2,760	325	527	2,713	298	14,201
2012	177	1,850	1,404	537	2,756	260	551	2,671	285	476	2,599	255	13,821
2013	194	1,627	1,403	688	2,649	242	676	2,493	292	745	2,422	279	13,710
2014	215	1,713	1,443	665	2,576	258	775	2,443	288	922	2,339	276	13,913
2015	188	1,739	1,545	713	2,671	253	850	2,496	278	802	2,527	261	14,323
2016	164	2,012	1,289	692	2,639	235	653	2,622	248	661	2,571	243	14,029
2017	153	1,763	1,490	549	2,689	231	583	2,613	230	660	2,528	224	13,713
2018	151	1,773	1,315	602	2,709	240	618	2,587	224	752	2,464	225	13,660
Total	5,136	30,376	43,587	14,458	72,576	8,789	12,180	70,577	9,465	12,906	69,546	8,536	358,132

## 3.2 Variable Measurement

### 3.2.1 Document Similarity

I measure the quarter-on-quarter similarities between 10K/10Q filings using the similarity measures from Cohen et al. (2020). Large (small) values of the similarity measures indicate small (large) changes in textual narratives between the two documents. In this study, I focus on the two commonly used similarity measures: Cosine similarity and Jaccard similarity.

(1) Cosine similarity, a measure of similarity between two non-zero vectors of an inner product space. The cosine similarity between two documents is defined as

$$Sim\_Cosine = \frac{D_1^{TF} \cdot D_2^{TF}}{\|D_1^{TF}\| \times \|D_2^{TF}\|} \quad (1)$$

Where  $D_1^{TF}$  is the term frequency vectors of document  $D_1$ ;  $D_2^{TF}$  is the term frequency vectors of document  $D_2$ ; The dot product is the scalar product and the norm, and  $\| \quad \|$  is the Euclidean norm.

(2) Jaccard similarity, a measure of similarity computed as the size of the intersection divided by the size of the union of the two term frequency sets. The Jaccard similarity between two documents is defined as

$$Sim\_Jaccard = \frac{|D_1^{TF} \cap D_2^{TF}|}{|D_1^{TF} \cup D_2^{TF}|} \quad (2)$$

Where  $D_1^{TF}$  is the term frequency vectors of document  $D_1$ ;  $D_2^{TF}$  is the term frequency vectors of document  $D_2$ ;  $\cap$  is the intersection;  $\cup$  is the union.



I cleaned all the 10K/10Q filings before I calculate the similarity scores. Detailed information about the data cleansing processes is in Section 3 of Appendix B Document Similarity and Latent Dirichlet Allocation.

### 3.2.2 Topic Overlap and Comparison Language

In this study I examine two text characteristics of the 10K/10Q filings and earnings conference call transcripts: the topic overlap and comparison language.

(1) Topic overlap, a measure captures the extent of overlap between the topics on conference call transcripts and those on 10K/10Q filings. Specifically, I apply the Latent Dirichlet Allocation (LDA) topic modeling method to both 10K/10Q filings and earnings conference call transcripts to extract topics for each document.<sup>7</sup> Based on the topics in each pair of 10K/10Q filing and earnings conference call transcripts, I calculate two topic overlap measures.<sup>8</sup>

The first topic overlap measure is the topic based Jaccard similarity (*TopicOverlap*). This measure is calculated using the equation below.

$$TopicOverlap = \frac{|\text{Set \{topics in conference call transcripts\} } \cap \text{Set \{topics in 10K/10Q filings\}}|}{|\text{Set \{topics in conference call transcripts\} } \cup \text{Set \{topics in 10K/10Q filings\}}|} \quad (3)$$

Where Set {topics in conference call transcripts} is the unique topics in a conference call transcript; Set {topics in 10K/10Q filings} is the unique topics in a

---

<sup>7</sup> The 10K/10Q filings used in the topic modeling are the same cleaned 10K/10Q filings that are used in the document similarity calculation. The earnings conference call transcripts are cleaned by using the same data cleansing processes as used in the 10K/10Q filings cleansing.

<sup>8</sup> As the topic overlap measures are based on the topics in the 10K/10Q filings and earnings conference call transcripts, the measures can not capture topics or information that are not disclosed in the 10K/10Q filings. Therefore, the study only examines the information that is disclosed in the 10K/10Q filings by the company but not the information that is concealed by the company.

10K/10Q filing; the numerator is the size of the intersection of the two sets; the denominator is the size of the union of the two sets.

The second topic overlap measure is the ratio of the number of overlapped topics in the 10K/10Q filings and earnings conference call transcripts to the number of topics in the 10K/10Q filings (*TopicCoverage*). This measure is calculated using the equation below.

$$TopicCoverage = \frac{|\text{Set \{topics in conference call transcripts\} } \cap \text{Set \{topics in 10K/10Q filings\}}|}{|\text{Set \{topics in 10K/10Q filings\}}|} \quad (4)$$

Where Set {topics in conference call transcripts} is the unique topics in a conference call transcript; Set {topics in 10K/10Q filings} is the unique topics in a 10K/10Q filing; the numerator is the size of the intersection of the two sets; the denominator is the size of Set {topics in 10K/10Q filings}.

For both topic overlap measures, a larger (small) value indicates more (less) topics overlap between earnings conference call transcripts and those on 10K/10Q filings.

Detailed information about the data cleansing, LDA models training and selection, and topics extraction and visualization are in Section 5 of Appendix B Document Similarity and Latent Dirichlet Allocation.

(2) Comparison language (*CompWord*), a measure captures the number of comparison words and phrases used on conference call transcripts. Like Cohen et al. (2020), I measure the comparison language by counting the number of a few comparison words and phrases in each earnings conference call transcript. The comparison words and phrases include “previous”, “compared”, “last year”, “prior year”, “previous year”,

“compared to” and “compared with.” Specifically, I calculate the comparison language measure using the equation below.

$$CompWord = \log\left(\frac{\text{Number of comparison words and phrases used in a conference call transcript}}{\text{length of the conference call transcript}} * 10000 + 1\right) \quad (5)$$

A larger (small) value of this measure indicates more (less) comparison language is used in the conference call transcript.

For each conference call, I calculate the comparison language measure based on the full conference call transcript. As the comparison language used in the Presentation session and that in the Question & Answer session of the conference call may have different effects on investors’ decision, I also separately calculate the comparison language measure based on the Presentation session and the Question & Answer session.

### 3.3 Method

#### 3.3.1 Confirmation of the Existence of Lazy Prices

Table IV of Cohen et al. (2020) proves the existent of lazy prices anomaly. This table reports results of Fama-Macbeth cross-sectional regressions of individual firm-level stock returns on the similarity measures and several return predictors. To test my hypothesis, I first confirm the existence of lazy prices by replicating the Table IV of Cohen et al. (2020). The model used in Cohen et al. (2020) is below.

$$Return = a_0 + a_1 Similarity + a_i Controls + \varepsilon \quad (6)$$

where *Return* is individual firm-level stock return in the following month after the 10K/10Q filing date, *Similarity* is one of the two similarity measures defined above, *Controls* are control variables define below.

The variable of primary interest is *Similarity*. A positive coefficient on *Similarity* replicates Cohen et al. (2020), indicating that larger changes in textual narratives (i.e., smaller values of *Similarity*) of financial reports are associated with *lower* future returns.

The control variables in the model include: *Size*, the log of the market value of equity;  $\log(BM)$ , the log of the book value of equity over market value of equity;  $Ret(-1, 0)$ , the previous month's return; and  $Ret(-12, -1)$ , the cumulative stock return from month -12 to month -1; *SUE*, the standardized unexpected earnings surprise.

### 3.3.2 Effect of Topic Overlap and Comparison Language on Lazy Prices

I test my hypotheses by using the equation below.

$$Return = a_0 + a_1 Similarity + a_2 CompWord(TopicOverlap, TopicCoverage) + a_3 Similarity * CompWord(TopicOverlap, TopicCoverage) + a_i Controls + \varepsilon \quad (7)$$

where  $CompWord(TopicOverlap, TopicCoverage)$  is the comparison language measure (topic overlap measures) defined above. All other variables are also defined above.

If the comparison language used in conference call transcripts and the topic overlap between 10K/10Q filings and conference call transcripts do mitigate lazy prices, I predict the coefficient on interaction between the interested measures with the similarity measure ( $a_3$ ) be significantly negative.

Appendix A defines all variables. All continuous variables are winsorized at the 1% and 99% levels to reduce the influence of outliers.

### 3.4 Validation of the Topic Modeling Approaches

LDA has some substantive disadvantages, one of which is nondeterministic, i.e., the LDA model generate different topics after each training even for the same corpus.

Therefore, without careful design, the results from LDA may be misleading. But extant studies rarely discuss the disadvantages. This study compares the topic coherence and topic stability of different LDA models in each industry to determine the preferred model in this study.<sup>9</sup> A model that can generate topics with high topic coherence and topic stability is preferred. This study also compares the performance of the Gensim's LdaMulticore model with that of the Tomotopy's LDA model by applying them to SEC 10K/10Q filings and earnings conference call. Finally, this study gives several suggestions to scholars who are interested in applying topic modeling methods to accounting research.

Detailed information about the LDA method is in Appendix B Document Similarity and Latent Dirichlet Allocation.

---

<sup>9</sup> Röder, Both and Hinneburg (2015) introduced and discussed the topic coherence measures.

## CHAPTER 4. EMPIRICAL RESULTS

### 4.1 Descriptive Statistics

Panel A of Table 4.1 presents the descriptive statistics of the main variables for the basic sample (1995-2018). The mean (median) one-month return ( $Ret$ ) after the release of the 10K/10Q filings is 0.905 (0.417). Each of the two similarity measures ranges from zero to one in theory. The mean of the Cosine similarity ( $Sim\_Cosine$ ) is 0.833 with a standard deviation of 0.176. The mean of the Jaccard similarity ( $Sim\_Jaccard$ ) is 0.673 with a standard deviation of 0.148. For the similarity measures, higher values indicate a higher degree of document similarity across years between the pair of 10K/10Q filings, while lower values indicate more changes across documents.

For the control variables, the mean (median) of  $Size$  is 5.972 (5.904) with a standard deviation of 2.045. The mean (median) of  $logBM$  is -0.732 (-0.647) with a standard deviation of 0.858. The mean of  $Ret(-1,0)$  is 0.011 and the mean of  $Ret(-12,-1)$  is 0.125. The mean (median) of  $SUE$  is 0.002 (0.001) with a standard deviation of 0.069.

Panel B of Table 4.1 presents the descriptive statistics of the main variables for the earnings conference call sample (2005-2018). There are six new variables.  $TopicOverlap$  is the Jaccard similarity based on the topics in the 10K/10Q filings and earnings conference call transcripts.  $TopicCoverage$  is the ratio of the number of overlapped topics in the 10K/10Q filings and earnings conference call transcripts to the number of topics in the 10K/10Q filings. For both the  $TopicOverlap$  and  $TopicCoverage$ , a larger value indicates that more topics in the 10K/10Q filings are discussed in the earnings conference calls.  $CompWord$  is the log value of the ratio of the number of comparative words and phrases used in the earnings conference call to the length of the earnings conference call transcript

multiplied by 10000. *CompWord\_QA* is the log value of the ratio of the number of comparative words and phrases used in the Question & Answer session of the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. *CompWord\_PR* is the log value of the ratio of the number of comparative words and phrases used in the Presentation session of the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. *Call\_Indicator* is a dummy variable which takes a value of one if the firm holds an earnings conference call, otherwise zero. The mean of *TopicOverlap (TopicCoverage)* is 0.110 (0.136) which means that, on average, an estimated 11%-13% of the topics in the 10K/10Q filings are discussed in the earnings conference calls. The mean of *CompWord* is 1.007 with a standard deviation of 1.063. The mean of *CompWord\_PR* (0.806) is larger than the mean of *CompWord\_QA* (0.563) which indicates that more comparative language is used in the Presentation session than the Q&A session of the earnings conference call. The mean of *Call\_Indicator* is 0.491 which means that 49.1% of the firms in the sample hold quarterly earnings conference calls.

Table 4.1 Descriptive Statistics

<b>Panel A: Descriptive Statistics of the Basic Sample (1995-2018)</b>						
Variable	N	Mean	Std	P25	Median	P75
<i>Ret</i>	358,132	0.905	13.677	-5.695	0.417	6.564
<i>Sim_Cosine</i>	358,132	0.833	0.176	0.769	0.905	0.959
<i>Sim_Jaccard</i>	358,132	0.673	0.148	0.580	0.691	0.786
<i>Size</i>	358,132	5.972	2.045	4.458	5.904	7.380
<i>logBM</i>	358,132	-0.732	0.858	-1.208	-0.647	-0.177
<i>Ret (-1,0)</i>	358,132	0.011	0.159	-0.070	0.004	0.081
<i>Ret (-12,-1)</i>	358,132	0.125	0.569	-0.205	0.056	0.328
<i>SUE</i>	358,132	0.002	0.069	-0.006	0.001	0.007
<b>Panel B: Descriptive Statistics of the Earnings Conference Call Sample (2005-2018)</b>						
Variable	N	Mean	Std	P25	Median	P75
<i>Ret</i>	199,911	0.729	12.343	-5.179	0.410	5.931
<i>Sim_Cosine</i>	199,911	0.879	0.145	0.851	0.941	0.972
<i>Sim_Jaccard</i>	199,911	0.729	0.126	0.656	0.750	0.824
<i>TopicOverlap</i>	199,911	0.110	0.136	0.000	0.000	0.200
<i>TopicCoverage</i>	199,911	0.136	0.167	0.000	0.000	0.250
<i>CompWord</i>	199,911	1.007	1.063	0.000	0.000	2.039
<i>CompWord_QA</i>	199,911	0.563	0.648	0.000	0.000	1.179
<i>CompWord_PR</i>	199,911	0.806	0.904	0.000	0.000	1.635
<i>Call_Indicator</i>	199,911	0.491	0.500	0.000	0.000	1.000
<i>Size</i>	199,911	6.399	2.003	4.956	6.373	7.792
<i>logBM</i>	199,911	-0.758	0.861	-1.238	-0.667	-0.190
<i>Ret (-1,0)</i>	199,911	0.007	0.149	-0.067	0.004	0.075
<i>Ret (-12,-1)</i>	199,911	0.103	0.502	-0.180	0.057	0.296
<i>SUE</i>	199,911	0.002	0.067	-0.005	0.001	0.007

## 4.2 Correlation Analysis

Table 4.2 presents the Pearson correlations among the main variables used in the earnings conference call sample (2005-2018). All the variables except *Ret (-12,-1)* have a significant positive correlation with return (*Ret*). The correlation between the cosine similarity (*Sim\_Cosine*) and the Jaccard similarity (*Sim\_Jaccard*) is 0.824 which means the two document similarity measures are highly correlated. The topic overlap measures (*TopicOverlap* and *TopicCoverage*), and comparison language (*CompWord*) are positively associated with the similarity measures (*Sim\_Cosine* and *Sim\_Jaccard*).



Table 4.2 Correlation Matrix for the Main Variables

		A	B	C	D	E	F	G	H	I	J	K	L
<i>Ret</i>	A	1.000											
<i>Sim_Cosine</i>	B	0.014***	1.000										
<i>Sim_Jaccard</i>	C	0.040***	0.824***	1.000									
<i>TopicOverlap</i>	D	0.017***	0.043***	0.083***	1.000								
<i>TopicCoverage</i>	E	0.016***	0.038***	0.078***	0.987***	1.000							
<i>CompWord</i>	F	0.022***	0.023***	0.088***	0.815***	0.821***	1.000						
<i>Call_Indicator</i>	G	0.022***	0.019***	0.086***	0.827***	0.831***	0.966***	1.000					
<i>Size</i>	H	0.013***	-0.053***	-0.013***	0.267***	0.261***	0.342***	0.392***	1.000				
<i>logBM</i>	I	0.012***	0.056***	0.024***	-0.075***	-0.077***	-0.087***	-0.104***	-0.319***	1.000			
<i>Ret (-1,0)</i>	J	0.104***	0.009***	0.021***	0.016***	0.014***	0.031***	0.029***	0.019***	0.046***	1.000		
<i>Ret (-12,-1)</i>	K	-0.003	-0.001	0.021***	0.017***	0.016***	0.040***	0.039***	0.183***	-0.280***	-0.001	1.000	
<i>SUE</i>	L	0.025***	-0.003	-0.004**	-0.014***	-0.014***	-0.008***	-0.011***	-0.029***	-0.061***	0.073***	0.101***	1.000

Note: \*, \*\*, and \*\*\* indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

### 4.3 Confirmation of the Existence of Lazy Prices

Table 4.3 reports the results of the confirmation of lazy prices. The results are based on Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures and several known return predictors.

The coefficients of *Sim\_Cosine* in columns (1)-(3) are all significantly positive which means that a larger change in document similarity is associated with a larger negative return per month in the future. The results hold even after controlling for the effects of several known predictors including *Size*, *logBM*, *Ret (-1,0)*, *Ret (-12,-1)*, *SUE*. Similarly, the coefficients of *Sim\_Jaccard* in columns (4)-(6) are all significantly positive. These results are consistent with those in Cohen et al. (2020). However, the coefficients of the similarity measures in the replication results are larger than those in Cohen et al. (2020). One reason is that Cohen et al. (2020) utilize the quintiles of the similarity measures based on the prior month's distribution of the similarity scores across all stocks while I utilize the raw value of the similarity measures. However, I find a similar pattern of the coefficient size of the similarity measures as in Cohen et al.(2020). The size of the coefficient on *Sim\_Jaccard* is twice the size of the coefficient on *Sim\_Cosine*.

For the control variables, I find a significant association between *Size* and *Return* while the association in Cohen et al.(2020) is insignificant. I find a significant association between *Ret (-1,0)* and *Return* while the association in Cohen et al. (2020) is significantly negative. Cohen et al. (2020) directly use the monthly return in their analysis while I calculated the monthly return based on daily returns. Because the dates of earnings conference calls and 10K/10Q filings releasing are important in the earnings conference

call analysis and the accurate future return relies on the starting date of the return calculation. I find similar results for *logBM*, *Ret (-12, -1)*, and *SUE*.

Table 4.3 Test of Lazy Prices Based on Cohen et al. (2020) Table IV (1995-2014)

This table reports results of Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures and several known return predictors. *Return*, the dependent variable, is the one-month return multiplied by 100 after the release of the 10K/10Q filing. *Sim\_Cosine*, is the cosine similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Sim\_Jaccard*, is the Jaccard similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Size* is the log of market value of equity. *log(BM)* is the log of book value of equity over market value of equity. *Ret(-1,0)* is the previous month's return, and *Ret(-12,-1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings. t-Statistics are reported below the estimates. Statistical significance at the 1%, 5%, and 10% levels is indicated by \*\*\*, \*\*, and \*, respectively.

	Dependent Variable: <i>Return</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Sim_Cosine</i>	0.869** (2.548)	1.070*** (3.318)	1.024*** (3.209)			
<i>Sim_Jaccard</i>				2.228*** (3.798)	2.369*** (4.411)	2.295*** (4.293)
<i>Size</i>		0.142*** (2.867)	0.145*** (2.948)		0.143*** (2.906)	0.146*** (2.985)
<i>logBM</i>		0.528*** (4.570)	0.546*** (4.732)		0.522*** (4.540)	0.540*** (4.700)
<i>Ret (-1,0)</i>		5.086*** (7.469)	4.807*** (7.092)		5.132*** (7.645)	4.852*** (7.273)
<i>Ret (-12,-1)</i>		-0.046 (-0.200)	-0.168 (-0.730)		-0.045 (-0.199)	-0.170 (-0.747)
<i>SUE</i>			5.410*** (3.753)			5.474*** (3.806)
<i>cons</i>	0.142 (0.303)	-0.820 (-1.595)	-0.754 (-1.492)	-0.601 (-1.055)	-1.516** (-2.558)	-1.440** (-2.454)
N	305,725	305,571	305,571	305,725	305,571	305,571
Avg. R <sup>2</sup>	0.01	0.05	0.06	0.01	0.06	0.07

Table 4.4 reports the results of Cohen et al. (2020) Table IV for the full sample (1995-2018). The results are similar with the replication results (1995-2014). The

coefficients of the document similarity measures further confirm the “lazy prices” in Cohen et al. (2020).

Table 4.4 Extension of Cohen et al. (2020) Table IV (1995-2018)

This table reports results of Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures and several known return predictors. *Return*, the dependent variable, is the one-month return multiplied by 100 after the release of the 10K/10Q filing. *Sim\_Cosine*, is the cosine similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Sim\_Jaccard*, is the Jaccard similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Size* is the log of market value of equity. *log(BM)* is the log of book value of equity over market value of equity. *Ret(-1,0)* is the previous month’s return, and *Ret(-12,-1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings. t-Statistics are reported below the estimates. Statistical significance at the 1%, 5%, and 10% levels is indicated by \*\*\*, \*\*, and \*, respectively.

	Dependent Variable: <i>Return</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Sim_Cosine</i>	0.857** (2.581)	1.211*** (4.042)	1.210*** (3.833)			
<i>Sim_Jaccard</i>				2.172*** (4.078)	2.465*** (5.148)	2.416*** (4.965)
<i>Size</i>		0.132*** (2.982)	0.124*** (2.680)		0.133*** (3.007)	0.125*** (2.678)
<i>logBM</i>		0.462*** (4.413)	0.435*** (3.759)		0.456*** (4.387)	0.432*** (3.802)
<i>Ret(-1,0)</i>		6.087*** (9.127)	5.775*** (9.072)		6.096*** (9.340)	5.774*** (9.288)
<i>Ret(-12,-1)</i>		-0.101 (-0.386)	-0.264 (-0.921)		-0.079 (-0.320)	-0.226 (-0.871)
<i>SUE</i>			10.781* (1.763)			10.703* (1.798)
<i>cons</i>	0.071 (0.165)	-0.928** (-1.993)	-0.892* (-1.952)	-0.669 (-1.305)	-1.590*** (-2.965)	-1.515*** (-2.868)
N	358,130	357,948	357,948	358,130	357,948	357,948
Avg. R <sup>2</sup>	0.01	0.06	0.07	0.01	0.06	0.07

#### 4.4 The Effect of Topic Overlap on Lazy Prices (H1)

To test whether the topic overlap between earnings conference calls and 10K/10Q filings can mitigate lazy prices, I interact the topic overlap variables with the document similarity variables. I expect a significantly negative coefficient for the interaction if the topic overlap can mitigate the lazy prices.

Table 4.5 shows the results. Column (1) and (2) presents the results based on the cosine similarity (*Sim\_Cosine*). The coefficients of *Sim\_Cosine\*TopicOverlap* and *Sim\_Cosine\*TopicCoverage* are all significantly negative at 5% level indicating that the topic overlap reduces the return predictive ability of the cosine similarity. Column (3) and (4) presents the results based on the Jaccard similarity (*Sim\_Jaccard*). Similarly, the coefficients of *Sim\_Jaccard\*TopicOverlap* and *Sim\_Jaccard\*TopicCoverage* are all significantly negative at 5% level. These results support my first hypothesis that topic overlap between earnings conference call transcripts and 10K/10Q filings mitigates lazy prices. Discussion of topics in the earnings conference call may help investors understand the topics including how the topics change compared to those in prior year and how the topics affect firm's future perform. Therefore, information related to the topics can be impounded into price timely. If these topics are also important topics in the 10K/10Q filings, they will not be able to predict the firm's future return when the 10K/10Q filings are released as they are already impounded into price before the release of 10K/10Q filings. Therefore, more topic overlap between earnings conference calls and 10K/10Q filings can mitigate lazy prices.

Table 4.5 The Effect of Topic Overlap on Lazy Prices

This table reports results of Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures, the comparison language measure, and several known return predictors. *Return*, the dependent variable, is the one-month return multiplied by 100 after the release of the 10K/10Q filing. *Sim\_Cosine*, is the cosine similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Sim\_Jaccard*, is the Jaccard similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *TopicOverlap* is the Jaccard similarity based on the topics in the 10K/10Q filings and earnings conference call transcripts, and a larger value indicates that the topics in the 10K/10Q filings and earnings conference call transcripts are more similar. *TopicCoverage* is the ratio of the number of overlapped topics in the 10K/10Q filings and earnings conference call transcripts to the number of topics in the 10K/10Q filings, and a larger value indicates that more topics in the 10K/10Q filings are discussed in the earnings conference calls. *Call\_Indicator* is a dummy variable which takes a value of one if the firm holds an earnings conference, otherwise zero. *Size* is log of market value of equity. *log(BM)* is the log of book value of equity over market value of equity. *Ret(-1,0)* is the previous month's return, and *Ret(-12,-1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings. t-Statistics are reported below the estimates. Statistical significance at the 1%, 5%, and 10% levels is indicated by \*\*\*, \*\*, and \*, respectively.

	Dependent Variable: <i>Return</i>			
	(1)	(2)	(3)	(4)
<i>Sim_Cosine</i>	1.757*** (2.959)	1.683*** (2.854)		
<i>Sim_Cosine*TopicOverlap</i>	-7.895** (-2.312)			
<i>Sim_Cosine*TopicCoverage</i>		-6.176** (-2.236)		
<i>Sim_Jaccard</i>			3.191*** (3.878)	3.059*** (3.751)
<i>Sim_Jaccard*TopicOverlap</i>			-11.917** (-2.550)	
<i>Sim_Jaccard*TopicCoverage</i>				-8.616** (-2.315)
<i>TopicOverlap</i>	7.965** (2.434)		9.702*** (2.661)	
<i>TopicCoverage</i>		6.255** (2.413)		7.057** (2.493)
<i>Call_Indicator</i>	0.212 (0.708)	0.239 (0.793)	0.182 (0.608)	0.215 (0.713)
<i>Size</i>	0.137** (2.118)	0.141** (2.219)	0.134** (2.026)	0.138** (2.121)
<i>logBM</i>	0.281* (1.665)	0.290* (1.729)	0.270 (1.579)	0.280* (1.656)
<i>Ret (-1,0)</i>	7.294*** (8.774)	7.276*** (8.801)	7.327*** (8.914)	7.312*** (8.949)

Table 4.5 (Continue)

	Dependent Variable: <i>Return</i>			
	(1)	(2)	(3)	(4)
<i>Ret (-12,-1)</i>	-0.538 (-1.353)	-0.562 (-1.388)	-0.545 (-1.424)	-0.577 (-1.452)
<i>SUE</i>	21.033 (1.275)	20.862 (1.279)	21.955 (1.275)	21.509 (1.281)
<i>cons</i>	-2.117*** (-2.818)	-2.101*** (-2.775)	-2.906*** (-3.426)	-2.854*** (-3.340)
N	199,911	199,911	199,911	199,911
Avg. R <sup>2</sup>	0.09	0.09	0.09	0.09

#### 4.5 The Effect of Comparison Language on Lazy Prices (H2)

To test whether the comparison language used in the earnings conference call can mitigate lazy prices, I interact the comparison language variable with the two document similarity variables. I expect a significantly negative coefficient for the interaction if the comparison language can mitigate the lazy prices.

Table 4.6 shows the results. Column (1) and (2) reports the results based on the full sample. In Column (1), the coefficient on *Sim\_Cosine\*CompWord* is insignificant, but the sign is negative. In Column (2), the coefficient on *Sim\_Jaccard\*CompWord* is negative and marginally significant at 10% level. The results are similar with those in Cohen et al.(2020) Table VIII. The coefficient on the interaction between the investor attention and document similarity is insignificant for the Cosine Similarity (*Sim\_Cosine*). But the coefficient for the Jaccard Similarity (*Sim\_Jaccard*) is significantly negative.

As I argue that the comparison language used in the earnings conference calls can attract investors' attention to firms' financial reports, the effect may be more significant if the earnings conference call date is closer to the 10K/10Q releasing date. Therefore, I

further limit the sample to only include the observations that the day-difference between earnings announcement date and 10K/10Q releasing date is within the range [-5,0]<sup>10</sup>. Column (3) and (4) reports the results based on this reduced sample. In Column (3), the coefficient on *Sim\_Cosine\*CompWord* is still insignificant. However, the coefficient on *Sim\_Jaccard\*CompWord* is significantly negative at 1% level in Column (4). The significance level increases from 10% to 1%.

Overall, the results support my second hypothesis, i.e., the use of comparison language on earnings conference call transcripts mitigates lazy prices.

Table 4.6 The Effect of Comparison Language on Lazy Prices

This table reports results of Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures, the comparison language measure, and several known return predictors. *Return*, the dependent variable, is the one-month return multiplied by 100 after the release of the 10K/10Q filing. *Sim\_Cosine*, is the cosine similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Sim\_Jaccard*, is the Jaccard similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *CompWord* is the log value of the ratio of the number of comparative words and phrases used in the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. *Call\_Indicator* is a dummy variable which takes a value of one if the firm holds an earnings conference, otherwise zero. *Size* is the log of market value of equity. *log(BM)* is the log of book value of equity over market value of equity. *Ret(-1,0)* is the previous month's return, and *Ret(-12,-1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings. t-Statistics are reported below the estimates. Statistical significance at the 1%, 5%, and 10% levels is indicated by \*\*\*, \*\*, and \*, respectively. The results in column (1) and (2) are based on the full sample; The results in column (3) and (4) are based on the sample that earnings announcement date is within 5 days prior to the financial report release date.

	Dependent Variable: <i>Return</i>			
	(1)	(2)	(3)	(4)
<i>Sim_Cosine</i>	1.381 (1.559)		1.271 (0.568)	
<i>Sim_Cosine*CompWord</i>	-1.790 (-1.127)		-1.318 (-0.902)	
<i>Sim_Jaccard</i>		5.361*** (3.046)		5.622*** (3.290)

<sup>10</sup> The earnings conference call is usually held on the same day as the earnings announcement, or, one day after.



Table 4.6 (Continue)

	Dependent Variable: <i>Return</i>			
	(1)	(2)	(3)	(4)
<i>Sim_Jaccard*CompWord</i>		-4.741*		-3.186***
		(-1.737)		(-3.018)
<i>CompWord</i>	1.674	3.713*	1.427	2.357**
	(1.126)	(1.719)	(1.045)	(2.518)
<i>Call_Indicator</i>	0.461	0.356	-0.780	-0.394
	(0.674)	(0.522)	(-0.773)	(-0.373)
<i>Size</i>	0.132**	0.146**	0.334***	0.358***
	(2.067)	(2.462)	(3.495)	(4.059)
<i>logBM</i>	0.287*	0.293*	0.509**	0.537***
	(1.742)	(1.904)	(2.430)	(2.826)
<i>Ret (-1,0)</i>	7.371***	7.178***	12.412***	12.088***
	(8.748)	(9.059)	(10.529)	(10.768)
<i>Ret (-12,-1)</i>	-0.744	-0.562	-1.609**	-1.357***
	(-1.410)	(-1.568)	(-2.311)	(-3.117)
<i>SUE</i>	16.706	14.475	18.030	13.061
	(1.368)	(1.479)	(1.206)	(1.347)
<i>cons</i>	-1.731*	-4.626***	-2.393	-5.677***
	(-1.677)	(-3.315)	(-1.014)	(-4.035)
N	199,911	199,911	110,781	110,781
Avg. R <sup>2</sup>	0.09	0.09	0.14	0.14

## CHAPTER 5. ADDITIONAL ANALYSIS

### 5.1 Comparison Language in the Presentation Session and Q&A Session

Prior literature find that the Question and Answer (Q&A) session of the earnings conference call is relatively more informative than the Presentation session (Matsumoto et al. 2011). The comparison language used in the Q&A session may have different effect on investors' attention compared to the comparison language used in the Presentation session. Therefore, I calculate the comparison language variable for the Q&A session and Presentation session separately and test how the comparison language in different sessions affect investors' attention. Table 5.1 shows the results.

Table 5.1 The Effect of Comparison Language on Lazy Prices (Q&A VS. Presentation)

This table reports results of Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures, the comparison language measure, and several known return predictors. *Return*, the dependent variable, is the one-month return multiplied by 100 after the release of the 10K/10Q filing. *Sim\_Cosine*, is the cosine similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Sim\_Jaccard*, is the Jaccard similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *CompWord\_QA* is the log value of the ratio of the number of comparative words and phrases used in the Question & Answer session of the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. *CompWord\_PR* is the log value of the ratio of the number of comparative words and phrases used in the Presentation session of the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. *Call\_Indicator* is a dummy variable which takes a value of one if the firm holds an earnings conference, otherwise zero. *Size* is log of market value of equity. *log(BM)* is the log of the book value of equity over market value of equity. *Ret (-1,0)* is the previous month's return, and *Ret (-12,-1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings. t-Statistics are reported below the estimates. Statistical significance at the 1%, 5%, and 10% levels is indicated by \*\*\*, \*\*, and \*, respectively. The results in column (1) and (2) are based on the full sample; The results in column (3) and (4) are based on the sample that earnings announcement date is within 5 days prior to the financial report release date.

	Dependent Variable: <i>Return</i>			
	(1)	(2)	(3)	(4)
<i>Sim_Cosine</i>	3.525*** (2.737)	3.111** (2.059)		

Table 5.1 (Continue)

	Dependent Variable: <i>Return</i>			
	(1)	(2)	(3)	(4)
<i>Sim_Cosine*CompWord_QA</i>	-3.163** (-2.372)			
<i>Sim_Cosine*CompWord_PR</i>		-3.171** (-2.287)		
<i>Sim_Jaccard</i>			4.939*** (3.222)	5.829*** (2.932)
<i>Sim_Jaccard*CompWord_QA</i>			-3.841*** (-2.801)	
<i>Sim_Jaccard*CompWord_PR</i>				-3.774** (-2.397)
<i>CompWord_QA</i>	3.231*** (2.621)	0.561 (1.367)	3.300*** (2.932)	0.806** (2.069)
<i>CompWord_PR</i>	0.023 (0.048)	2.955** (2.160)	-0.054 (-0.124)	2.851** (2.213)
<i>Call_Indicator</i>	-0.503 (-0.477)	-0.910 (-1.009)	-0.498 (-0.488)	-1.255 (-1.270)
<i>Size</i>	0.331*** (3.474)	0.317*** (3.424)	0.343*** (3.659)	0.337*** (3.779)
<i>logBM</i>	0.504** (2.184)	0.504** (2.148)	0.484** (2.144)	0.484** (2.008)
<i>Ret (-1,0)</i>	11.634*** (10.031)	11.322*** (8.370)	11.671*** (10.074)	11.566*** (9.453)
<i>Ret (-12,-1)</i>	-1.340*** (-2.674)	-1.059*** (-2.885)	-1.491*** (-2.839)	-1.314*** (-3.041)
<i>SUE</i>	23.547 (1.141)	23.771 (1.132)	22.721 (1.158)	22.524 (1.165)
<i>cons</i>	-4.631*** (-3.434)	-4.284*** (-2.803)	-5.205*** (-3.683)	-5.890*** (-3.483)
N	110,781	110,781	110,781	110,781
Avg. R <sup>2</sup>	0.15	0.15	0.15	0.15

Column (1) and (2) presents the results based on the Cosine similarity (*Sim\_Cosine*). The coefficients on *Sim\_Cosine\*CompWord\_QA* and *Sim\_Cosine\*CompWord\_PR* are all significantly negative at 5% level. Column (3) and (4) presents the results based on the Jaccard similarity (*Sim\_Jaccard*). The coefficient on

*Sim\_Jaccard\*CompWord\_QA* is significantly negative at 1% level, and the coefficient on *Sim\_Jaccard\*CompWord\_PR* is significantly negative at 5% level. The results indicate that the comparison language used in both the Presentation session and the Q&A session can attract investors' attention to the firm's financial reports and thus mitigate lazy prices. Overall, the results support my second hypothesis that comparison language used in the conference call mitigates lazy prices.

## 5.2 Document Similarity based on Word Stem

Before calculating the document similarity, and training the LDA models, I tokenize the documents and transform the word to its stem form or lemma form.<sup>11</sup> It is much faster to transform the words to its stem than to its lemma. However, using the stem form of the words has its limitation that the word may lose its true meanings. For example, the stem of "studies" is "studi" while the lemmas of "studies" and "studying" are "study." It is more reasonable to use the lemma form of the words when analyzing the financial reports though it may dramatically increase the computing time. I use the lemma form of the words in all my analysis. However, I also test the effect of document similarity on future return by using the stem form of the words.

---

<sup>11</sup> "Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation." (Manning, Raghavan, and Schütze, 2008) (<https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html>)

"The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form". "Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma" (Manning, Raghavan, and Schütze, 2008) (<https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>). For example, the stem of "studies" is "studi" while the lemmas of "studies" and "studying" are "study".

Table 5.2 shows the results. The results are like those in the test that uses the lemma form of the words. The coefficients on Cosine Similarity (*Sim\_Cosine*) and Jaccard Similarity (*Sim\_Jaccard*) are all significantly negative. As I did not use the stem form of the words in the LDA modeling, I do not have results of the earnings conference call analysis based on the stem form of the words.

The results indicate that while the lemma form of words makes more sense when analyzing the financial reports, the stem form of words is also an alternative to the lemma form as its computing speed is much faster than that of the lemma form.

Table 5.2 The Effect of Document Similarity on a Firm's Future Return (Stem)

This table reports results of Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on two similarity measures and several known return predictors. *Return*, the dependent variable, is the one-month return multiplied by 100 after the release of the 10K/10Q filing. *Sim\_Cosine*, is the cosine similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Sim\_Jaccard*, is the Jaccard similarity measure which captures the quarter-over-quarter similarities between 10K/10Q filings, and a larger value indicates that the two documents are more similar. *Size* is log of market value of equity. *log(BM)* is the log of book value of equity over market value of equity. *Ret(-1,0)* is the previous month's return, and *Ret(-12,-1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings. t-Statistics are reported below the estimates. The document similarity measures are calculated by using the stem form of the words. Statistical significance at the 1%, 5%, and 10% levels is indicated by \*\*\*, \*\*, and \*, respectively.

	Dependent Variable: <i>Return</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Sim_Cosine</i>	1.058*** (2.730)	1.241*** (3.388)	1.182*** (3.255)			
<i>Sim_Jaccard</i>				2.713*** (3.927)	2.840*** (4.468)	2.751*** (4.352)
<i>Size</i>		0.142*** (2.874)	0.146*** (2.951)		0.145*** (2.955)	0.148*** (3.035)
<i>logBM</i>		0.531*** (4.595)	0.548*** (4.752)		0.532*** (4.620)	0.549*** (4.777)
<i>Ret (-1,0)</i>		5.089*** (7.487)	4.813*** (7.115)		5.105*** (7.592)	4.822*** (7.212)
<i>Ret (-12,-1)</i>		-0.043 (-0.187)	-0.164 (-0.716)		-0.037 (-0.164)	-0.161 (-0.704)
<i>SUE</i>			5.344*** (3.708)			5.388*** (3.755)
<i>cons</i>	0.019 (0.039)	-0.924* (-1.729)	-0.849 (-1.613)	-0.744 (-1.275)	-1.647*** (-2.725)	-1.567*** (-2.618)
N	305,725	305,571	305,571	305,725	305,571	305,571
Avg. R <sup>2</sup>	0.01	0.05	0.06	0.01	0.06	0.07

## CHAPTER 6. CONCLUSION

Changes to the language and construction of financial reports are indicative of firms' future returns and operations. Investors, however, are often inattentive to these changes; consequently, price reactions to these changes are delayed—resulting in “lazy” prices (Cohen et al., 2020). This study explores two possible channels through which earnings conference calls may mitigate lazy prices: (1) the topic overlap channel and (2) the comparison language channel. I find evidence that both channels work. Specifically, I find that the more topics overlap between conference call transcripts and 10K/10Q filings, and the more comparison language used on earnings conference call transcripts, the less “lazy” are the prices.

This study also has its limitation. This study applied LDA to financial reports and earnings conference call transcripts to extract topics. However, LDA is a non-deterministic topic model, and the output topics differs after each run. I did not compare the quality of the output topics with other deterministic topic models, such as principal component analysis (PCA) and non-negative Matrix Factorization (NMF), because of the availability of computing resources. To mitigate this limitation, this study carefully chose the preferred LDA models by comparing the topic coherence and topic stability among different models. This study also utilized the Tomotopy library to apply the LDA model. Tomotopy is a newly developed python library which focuses on topic modeling. Tomotopy can generate more accurate and stable topics compared to other topic modeling library.

Overall, this study contributes to both the earnings conference call literature and the capital market literature by showing that earnings conference calls can mitigate lazy prices through the “topic overlap channel” and the “comparison language channel.” This

study also contributes to the textual analysis literature in accounting by providing best practices of applying topic modeling methods to accounting research.



## APPENDICES

### APPENDIX 1. VARIABLE DEFINITIONS

Variable	Definition
<b>Main Variables</b>	
<i>Return</i>	The one-month return multiplied by 100 after the release of the 10K/10Q filing.
<i>Sim_Cosine</i>	The Cosine similarity measure used in Cohen et al. (2020) . This measure captures the quarter-over-quarter textual narratives similarities between 10K/10Q filings. A larger value indicates that the two documents are more similar.
<i>Sim_Jaccard</i>	The Jaccard similarity measure used in Cohen et al. (2020) . This measure captures the quarter-over-quarter textual narratives similarities between 10K/10Q filings. A larger value indicates that the two documents are more similar.
<i>TopicOverlap</i>	The Jaccard similarity based on the topics in the 10K/10Q filings and earnings conference call transcripts. A larger value indicates that the topics in the 10K/10Q filings and earnings conference call transcripts are more similar.
<i>TopicCoverage</i>	The ratio of the number of overlapped topics in the 10K/10Q filings and earnings conference call transcripts to the number of topics in the 10K/10Q filings. A larger value indicates that more topics in the 10K/10Q filings are discussed in the earnings conference calls.
<i>CompWord</i>	The log value of the ratio of the number of comparative words and phrases used in the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. A larger value indicates that more comparative words and phrases are used in the earnings conference call.
<i>CompWord_QA</i>	The log value of the ratio of the number of comparative words and phrases used in the Question & Answer session of the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. A larger value indicates that more comparative words and phrases are used in the Question & Answer session of the earnings conference call.
<i>CompWord_PR</i>	The log value of the ratio of the number of comparative words and phrases used in the Presentation session of the earnings conference call to the length of the earnings conference call transcript multiplied by 10000. A larger value indicates that more comparative words and phrases are used in the Presentation session of the earnings conference call.
<i>Call_Indicator</i>	Call_Indicator is a dummy variable which takes a value of one if the firm holds an earnings conference call, otherwise zero.

APPENDIX 1 (Continue)

Variable	Definition
<b>Control Variables</b>	
<i>Size</i>	Logarithm of the market value of equity. COMPUSTAT: $cshoq * prccq$
<i>logBM</i>	Logarithm of the book value of equity scaled by market value of equity at the end of the quarter. COMPUSTAT: $ceqq / (cshoq * prccq)$
<i>Ret (-1,0)</i>	The previous month's return.
<i>Ret (-12,-1)</i>	The cumulative return from month -12 to month -1.
<i>SUE</i>	The quarterly standardized earnings surprises based on time-series (seasonal random walk model) and exclude special items using methodology in Livnat and Mendenhall (JAR, 2006).

## APPENDIX 2. DOCUMENT SIMILARITY AND LATENT DIRICHLET ALLOCATION

### 1. Introduction

In this Appendix, I describe the main processes of calculating the document similarity for and applying the Latent Dirichlet Allocation (LDA) topic modeling method to the SEC 10K/10Q filings and earnings conference call transcripts. Specifically, I introduce how I prepare the data for the document similarity calculation and the LDA models. For document similarity, I introduce the main processes and the computer runtime. For the LDA topic modeling, I introduce how I create the corpus, train the LDA models, evaluate and select the LDA models, and apply the LDA models. I also present the main techniques in each process and the computer runtime to give scholars a benchmark of applying the LDA topic modeling method to SEC 10K/10Q filings. Finally, I compare the performance of the Tomotopy's LDA model with that of the Gensim's LdaMulticore model in terms of the model training time.

I utilize Python (version 3.6) to clean the data, calculate the document similarity, and apply the LDA models. The main Python libraries utilized in this study include but are not limited to NLTK, Gensim, Tomotopy, Pandas, Scikit-learn, and Matplotlib.

This appendix proceeds as follows. Section 2 introduces the machines I utilized for this study. Section 3 explains the data preparation process. Section 4 describes the document similarity calculation. Section 5 presents the main processes of the LDA modeling method. Section 6 compares different Topic Modeling methods. Section 7 concludes the study with a discussion of the suggestions and limitations.

## 2. Machines

This study applied Natural Language Processing (NLP) techniques. NLP requires high computing power when processing a large corpus. Hence, I described the machines I used in this study (data cleansing, document similarity calculation, and LDA modeling). This helps scholars who plan to apply the methods used in this study to get a benchmark for their own research projects.

Table A1 shows the detailed machine information. Some of the NLP processes require high memory while other NLP processes require more processors to speed up the computing and save project time. Therefore, it is important to know the number of processors and the volume of the memory for each machine. The CPU type and base speed also affect the processing speed. A higher generation of the CPU and a higher base speed are preferred. The Surface Pro machine only has 4 processors and 16GB memory while the Dell Desktop machine has 8 processors and 16GB memory. The Dell Desktop machine also has a higher CPU generation and higher base speed compared to the Surface Pro. Therefore, the Dell Desktop machine has a higher computing power than the Surface Pro machine. The Windows Virtual Machine (Windows VM) and Linux Virtual Machine (Linux VM) have relatively higher computing power than the Surface Pro machine and Dell Desktop machine. The Windows VM has 32 processors and 128GB memory while the Linux VM has 80 processors and 1024GB memory. Therefore, the Linux VM is the most powerful machine. Both the Windows VM and Linux VM are from the cloud computing platform OpenStack through University of Kentucky. Table A1 shows how the Linux VM's performance exceeds other machine's performance in training the LDA models.

Table A1 Machine Information

Machine	System	CPU	# of Processors	Memory (GB)	Base Speed (GHz)
Surface Pro	Windows	Intel(R) Core™ i5-7300U	4	16	2.7
Dell Desktop	Windows	Intel(R) Core™ i7-6700	8	16	3.4
Windows Virtual Machine (Windows VM)	Windows	Intel Core (Broadwell, IBRS)	32	128	2
Linux Virtual Machine (Linux VM)	Linux	Intel Core (Broadwell, IBRS)	80	1024	NA

### 3. Data Preparation

Data preparation is the most important process before calculating the document similarity and applying the LDA models. The quality of the data preparation determines the quality of the output. This study cleaned 921,265 SEC 10K/10Q filings (1994-2018) and 152,187 earnings conference call transcripts (2005-2018).

The data cleansing processes include (1) removing html tags, (2) converting all words to lower case, (3) tokenizing the document, (4) removing stop words, numbers, and words that are only one character, and (5) lemmatizing the tokens in the document.<sup>12</sup> To save storage space, all the tokenized documents were encoded and compressed by using base64 and zlib. Table A2 provides information about the first round of data cleansing.

<sup>12</sup> “Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.”(Manning, Raghavan, and Schütze, 2008) (<https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html>)

“The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.” “Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.”(Manning, Raghavan, and Schütze, 2008) (<https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>). For example, the stem of “studies” is “studi” while the lemmas of “studies” and “studying” are “study”.

In the first round of data cleansing, I cleaned 626,304 documents including 474,428 10K/10Q filings and 151,876 earnings conference call transcripts. The total size of the 626,304 documents is nearly 90GB. The lemmatization process was extremely time-consuming and required high memory. For a single processor, it took one minute to clean four documents on average (the earnings conference call transcript takes relatively less time than the SEC 10K/10Q filing as the transcript is relatively shorter). Cleaning the 626,304 documents required 2,610 hours for a single CPU processor. In this round, I utilized all the machines simultaneously to do the data cleansing as I only got access to the two virtual machines in a later stage. Taking advantage of the multiple processors in each machine, I spent a total of 340.5 hours to clean the 626,304 documents.

Table A2 First Round of Data Cleansing Information

Machine	# of Processors	Average Files/minute	File Type	Total Files	Total Hours
Surface Pro	4	15	SEC 10K/10Q, Call Transcripts	50,000	55
Dell Desktop	8	25	SEC 10K/10Q, Call Transcripts	336,304	224
Windows VM	32	20	SEC 10K/10Q	70,000	58
Linux VM	80	800	SEC 10K/10Q	170,000	3.5
Total			SEC 10K/10Q, Call Transcripts	626,304	340.5

Note: The total 626,304 files include 474,428 10K/10Q filings and 151,876 earnings conference call transcripts. The data cleansing processes include (1) removing html tags, (2) converting all the words to lower case, (3) tokenizing the document, (4) removing stop words, numbers, and words that are only one character, and (5) lemmatizing the tokens in the document.

To calculate the document similarity of the SEC 10K/10Q filings, I cleaned all the 921,265 SEC 10K/10Q filings (1994-2018). Therefore, I utilized the Linux VM to do the second-round of data cleansing. After the two rounds of data cleansing, the total size of the 921,265 cleaned and compressed SEC 10K/10Q documents is 17.5 GB and the total size of the 151,876 cleaned and compressed earnings conference call transcripts is 1.4 GB.

Table A3 shows the data cleansing information for this round. As Linux VM is powerful, it required only a total of 15 hours to clean the data.

Based on the information of the first round and second round of data cleansing, I recommend that scholars utilize the cloud computing services (e.g., Amazon Web Services and OpenStack) to get access to a virtual machine with a high memory and more processors in their “big text data” analysis.

Table A3 Second Round of Data Cleansing Information

Machine	# of Processors	File Type	Total Files	Total Hours
Linux VM	80	SEC 10K/10Q (1994-2004)	446,837	5.5
Linux VM	80	SEC 10K/10Q (2005-2018)	474,428	9.5
Total	80	SEC 10K/10Q (1994-2018)	921,265	15

Note: The data cleansing processes include (1) removing html tags, (2) converting all the words to lower case, (3) tokenizing the document, (4) removing stop words, numbers, and words that are only one character, and (5) lemmatizing the tokens in the document. To save storage space, all the tokenized documents are encoded and compressed by using base64 and zlib.

#### 4. Document Similarity

The document similarity captures the quarter-over-quarter textual narratives similarities between 10K/10Q filings. The main processes to obtain the document similarity scores include data cleansing and document similarity calculation. Figure 1 shows the document similarity calculation processes.

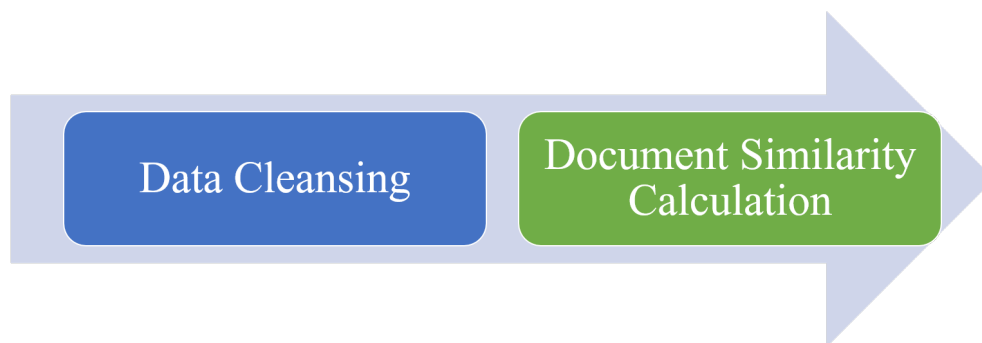


Figure A1 Document Similarity Calculation Processes

As introduced in the Data Preparation session, I had already cleaned the 921,265 SEC 10K/10Q filings (1994-2018). Next, I matched each 10K/10Q document with the document in previous year for each firm. For example, I matched the 10K (10Q) document of Apple Inc. (CIK # 0000320193) in year 2015 (quarter 3, 2015) with the 10K (10Q) document of Apple Inc. in year 2014 (quarter 3, 2014). Then I calculated the document similarity for each pair of 10K/10Q documents. Cosine similarity and Jaccard similarity are commonly utilized methods to calculate document similarity. I utilized Python to do the calculation. Specifically, I utilized the TfidfVectorizer from the sklearn.feature\_extraction.text in the scikit-learn library to vectorize the document to calculate the Cosine similarity. The calculation was relatively faster compared to the data cleansing. Table A4 shows the calculation time information. It only took 8.75 hours to calculate the Cosine similarity and Jaccard similarity for the 766,922 pairs of 10K/10Q filings by using the Linux virtual machine, which has 80 processors and 1024GB RAM memory.

Table A4 Document Similarity Calculation Information

Machine	# of Processors	File Type	Total Pairs	Total Hours
Linux VM	80	SEC 10K/10Q (1994-2018)	766,922	8.75

Note: The document similarity measures include Cosine Similarity and Jaccard Similarity.

## 5. Latent Dirichlet Allocation (LDA)

There are many topic modeling methods. Latent Dirichlet Allocation (LDA) is one of the most frequently utilized topic modeling methods, developed in Blei, Ng, and Jordan (2003). Several Python libraries that support LDA include Scikit-learn, Gensim, and



Tomotopy.<sup>13</sup> In this study, I utilized the Tomotopy library (version 0.10.1) to do the LDA analysis. Compared to the gensim's LdaModel, the LDA models in the Tomotopy are faster and more accurate.<sup>14</sup> I provided comparison results regarding the performance between the two libraries in Section 6. However, one of the disadvantages of Tomotopy is that it requires high memory to perform the LDA. The gensim's LdaModel requires less memory.

To ensure the reliability and stability of the LDA model, I trained the models by using as many documents as possible. Specifically, the corpus used to train the LDA models includes 474,428 10K/10Q filings and 151,876 earnings conference call transcripts. As the corpus is very large, the parsing process and the training process are very time-consuming. Below are the main steps to apply the LDA model to generate topics.

In this study, I utilized 469,918 SEC filings and 107,413 earnings conference call transcripts to train the LDA models. Specifically, I divided the sample into 12 subsamples based on the Fama-French 12 industry.<sup>15</sup> I chose the Fama-French 12 industry for three reasons. First, the 12 industries can ensure enough documents in each industry to train the LDA model and thus ensure the reliability and stability of the result. Second, including a

---

<sup>13</sup> The scikit-learn library: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>;

The Gensim library: <https://radimrehurek.com/gensim/models/ldamodel.html>

The Tomotopy library: <https://bab2min.github.io/tomotopy/v0.11.1/en/>

<sup>14</sup> Tomotopy is relatively new, and its first version is released on 2019-05-12. According to its official website, "tomotopy is a Python extension of tomoto (Topic Modeling Tool) which is a Gibbs-sampling based topic model library written in C++. It utilizes a vectorization of modern CPUs for maximizing speed." Regarding the performance, Tomotopy's official website provides a comparison with the gensim's LdaModel. The following is the statement.

"Tomotopy uses Collapsed Gibbs-Sampling(CGS) to infer the distribution of topics and the distribution of words. Generally, CGS converges more slowly than Variational Bayes(VB) that gensim's LdaModel uses, but its iteration can be computed much faster. In addition, tomotopy can take advantage of multicore CPUs with a SIMD instruction set, which can result in faster iterations." (Tomotopy: <https://bab2min.github.io/tomotopy/v0.11.1/en/>)

Though the Gensim' LDA model also allows multicore training, its performance is still relatively slower than the Tomotopy's LDA model.

<sup>15</sup> Website: [https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data\\_Library/det\\_12\\_ind\\_port.html](https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_12_ind_port.html)

reasonable number of documents in each industry can ensure the computer memory is enough to train the Tomotopy LDA model.<sup>16</sup> Third, topic extraction based on industry makes sense. For example, Huang, Leheavy, Zang, and Zheng Rong (2018) analyzed analyst reports and earnings conference call transcripts based on 4-digit Global Industry Classification Standard (GICS) industry classification.

Table A5 shows the 12 industries and the document distribution among the 12 industries.

Table A5 Document Distribution Base on Fama-French 12 Industry

Industry Number	# of Documents	Percent	Industry Name	Industry Detail
1	23,445	4.06%	Nondurables	Consumer Nondurables (Food, Tobacco, Textiles, Apparel, Leather, Toys)
2	12,241	2.12%	Durables	Consumer Durables (Cars, TVs, Furniture, Household Appliances)
3	40,578	7.03%	Manufacturing	Manufacturing (Machinery, Trucks, Planes, Off Furn, Paper, Com Printing)
4	28,307	4.90%	Enrgy	Oil, Gas, and Coal Extraction and Products
5	13,988	2.42%	Chemicals	Chemicals and Allied Products
6	78,683	13.63%	Business Equipment	Business Equipment (Computers, Software, and Electronic Equipment)
7	14,649	2.54%	Telecom	Telephone and Television Transmission
8	19,261	3.34%	Utilities	Utilities
9	47,339	8.20%	Shops	Wholesale, Retail, and Some Services (Laundries, Repair Shops)
10	58,244	10.09%	Healthcare	Healthcare, Medical Equipment, and Drugs
11	145,157	25.14%	Money	Finance
12	95,439	16.53%	Other	Other (Mines, Constr, BldMt, Trans, Hotels, Bus Serv, Entertainment)
Total	577,331	100.00%		

Note: The sample used to train the LDA model include 469, 918 SEC 10K/10Q filings (2005-2018) and 107, 413 matched earnings conference call transcripts. SIC industry code is used to classify the Fama-French 12 industries.

<sup>16</sup> For example, the Surface Pro machine with 16GB memory will fail to process a 10K/10Q corpus with more than 50, 000 documents due to the lack of memory problem.

The main processes of training and applying the LDA model includes data cleansing, dictionary, and corpus creation, LDA model training, and LDA model application. Figure 2 shows the main processes. I have already introduced the data cleansing process and cleaned all the documents used in the LDA model. Next, I will introduce the other three processes in the following sections.

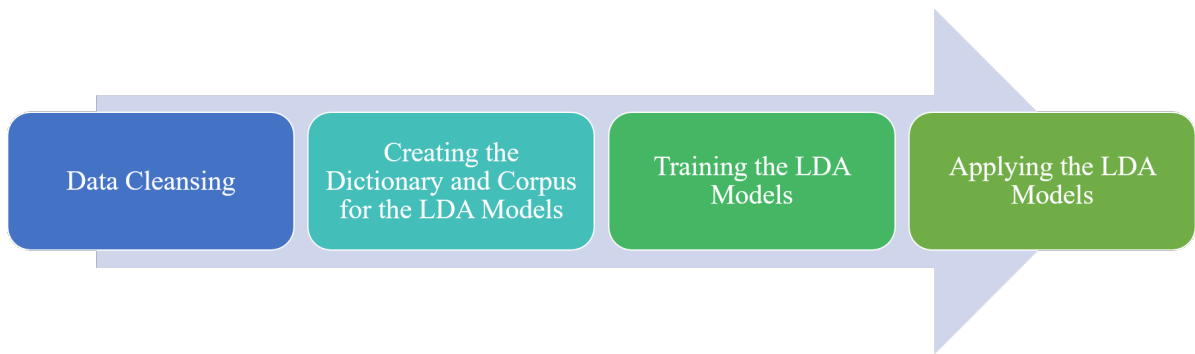


Figure A2 LDA Training and Applying Processes

### 5.1 Corpus Creation

Before training the LDA model, I need to create a corpus for each industry to manage the documents. I first created a dictionary based on all the words in all the documents and then filtered the extreme words, including words only in a few documents and words with very high frequency. Then, I utilized the dictionary to create a corpus for each industry. Table A6 presents the time used to create the corpus for each industry. To save time, all the machines were used simultaneously. The total computing time for creating the corporuses for the 12 industries is 227 minutes.

Table A6 Corpus Creation for LDA Models (Fama-French 12 Industry)

Machine	Industry	Documents	Total Time (minutes)
Surface Pro	1	23,445	29
Surface Pro	2	12,241	15
Dell Desktop	3	40,578	20
Dell Desktop	4	28,307	14
Surface Pro	5	13,988	22
Widows VM	6	78,683	38
Linux VM	7	14,649	1
Dell Desktop	8	19,261	14
Widows VM	9	47,339	23
Widows VM	10	58,244	30
Linux VM	11	145,157	13
Linux VM	12	95,439	8
Total		577,331	227

## 5.2 LDA Models Training

Training the LDA model using the Tomotopy library is fast but also memory intensive. For example, a machine with a 16GB RAM may fail to run the LDA model with a very large corpus (e.g., 50,000 10K/10Q documents). Therefore, I assigned the twelve corpuses to my four machines based on the corpus size. A smaller corpus was assigned to a machine with a lower RAM.

Before training the LDA models, it is important to decide a few hyperparameters and the number of topics for each model. In LDA a document is considered a probability distribution of topics and a topic a distribution over the words. Alpha is the hyperparameter of Dirichlet distribution for document-topic while eta is the hyperparameter of Dirichlet distribution for topic-word. I assign 0.1 for the alpha and 0.01 for the eta as they are commonly used. For the number of topics, I tried a range of numbers (30, 40, 45, 50, 60) based on prior research (Ball, Hoberg and Maksimovic, 2015; Dyer, Lang, and Stice-Lawrence, 2017; Huang, Lehavy, Zang and Rong, 2018; Brown, Crowley, and Elliott, 2020). For each industry, I trained 5 LDA models with different numbers of topics. Later

I chose the ideal number of topics by evaluating each model in each industry. The model with the ideal number of topics in each industry was used in this study. Table A7 shows the time information of training the LDA model with the ideal number of topics for each industry. This provides scholars a benchmark of training LDA models on SEC 10K/10Q filings and earnings conference call transcripts. The results show that it took about 16.5 hours (994 minutes) to train the 12 LDA models. As I trained 5 LDA models for each industry, the total number of LDA models that I trained was 60. It took about 79 hours (4,758 minutes) in total to train the 60 models. One advantage of the Tomotopy LDA model is that it takes advantage of multiple processors. Using a machine with 8 processor may take 10 times the time than using the Linux VM which has 80 processors to train the same model.

Table A7 LDA Models Training Information (Fama-French 12 Industry)

Machine	Industry	Documents	Min_df	Rm_top	# of Topics	Total Time (minutes)
Surface Pro	1	23,445	10	50	45	117
Surface Pro	2	12,241	10	50	40	50
Dell Desktop	3	40,578	20	50	60	108
Dell Desktop	4	28,307	20	50	45	80
Surface Pro	5	13,988	10	50	45	87
Widows VM	6	78,683	20	100	50	111
Linux VM	7	14,649	10	50	60	16
Dell Desktop	8	19,261	10	50	60	126
Widows VM	9	47,339	20	50	45	61
Widows VM	10	58,244	20	50	30	89
Linux VM	11	145,157	20	100	30	89
Linux VM	12	95,439	20	100	45	60
Total		577,331				994

Note: (1) This table only shows the training time of the LDA model with the ideal number of topics for each industry. The processes used to select the ideal number of topics are discussed in the LDA models evaluation and selection section. (2) The ideal number of topics is chosen

---

from a list [30,40,45,50,60]. Therefore, I trained 5 models with different number of topics for each industry. The total training time is 4,758 minutes. (3) The Min\_df and Rm\_top is used to filter the vocabularies used in the LDA model training. The number 10 in the Min\_df means that a word should appear in at least 10 documents; the number 50 in the Rm\_top means that the top 50 words are removed based on the frequency in the entire sample in each industry.

### 5.3 LDA Models Evaluation and Selection

After training all the LDA models for each industry, I evaluated the topic coherence and stability of each model to decide the best model to use in this study. In general, preferred models generate topics with high topic coherence and topic stability.

The module `tomotopy.coherence` provides a way to calculate the topic coherence introduced by Röder, Both, and Hinneburg (2015). The topic coherence measures include `u_mass`, `c_uci`, `c_npmi`, and `c_v`. In general, the value of `u_mass` measure ranges from -14 to 14 and a value that is close to zero is preferred. The values of `c_uci`, `c_npmi`, and `c_v` ranges from 0 to 1 and a larger value is preferred. I utilized the Jaccard similarity to measure the topic stability. The topic stability means less topic word overlap among topics in each model. Specifically, I calculated the mean Jaccard similarity of the topic words among topics for each LDA model. A smaller value of the Jaccard similarity means the model is more stable. In this study, I utilized the topic coherence measure `c_uci` and the topic stability measure Jaccard similarity to select the ideal LDA model. Table A8 presents the coherence score and stability score of each model in each industry. The Jaccard similarity scores of all models are less than 0.05 which indicates that the topic stability is generally high for the LDA models generated by the Tomotopy library.

Table A8 LDA Models Evaluation and Selection

Industry	# of Topics	Coherence Score			Stability Score
		u <sub>mass</sub>	c <sub>uci</sub>	c <sub>npmi</sub>	Jaccard
1	30	-0.709	0.649	0.099	0.023
1	40	-0.652	0.687	0.103	0.016
1	45	-0.676	0.743	0.105	0.020
1	50	-0.762	0.600	0.095	0.019
1	60	-0.954	0.575	0.096	0.017
2	30	-0.653	0.381	0.070	0.035
2	40	-0.627	0.425	0.070	0.032
2	45	-0.855	0.404	0.075	0.026
2	50	-0.651	0.406	0.074	0.029
2	60	-0.831	0.360	0.068	0.029
3	30	-0.607	0.609	0.084	0.028
3	40	-0.711	0.621	0.094	0.022
3	45	-0.674	0.576	0.086	0.027
3	50	-0.758	0.480	0.076	0.025
3	60	-0.644	0.631	0.088	0.026
4	30	-0.358	0.566	0.083	0.029
4	40	-0.459	0.574	0.088	0.026
4	45	-0.402	0.672	0.096	0.024
4	50	-0.566	0.663	0.097	0.024
4	60	-0.510	0.528	0.082	0.024
5	30	-0.541	0.517	0.080	0.024
5	40	-0.672	0.530	0.087	0.022
5	45	-0.579	0.530	0.085	0.024
5	50	-0.823	0.399	0.077	0.022
5	60	-0.742	0.460	0.079	0.027
6	30	-0.542	0.879	0.109	0.023
6	40	-0.638	0.806	0.106	0.017
6	45	-0.658	0.752	0.102	0.020
6	50	-0.659	0.922	0.111	0.020
6	60	-0.717	0.694	0.096	0.019
7	30	-0.413	0.784	0.106	0.023
7	40	-0.491	0.722	0.096	0.027
7	45	-0.583	0.741	0.104	0.023
7	50	-0.519	0.670	0.099	0.023
7	60	-0.559	0.789	0.107	0.022
8	30	-0.777	0.312	0.083	0.021
8	40	-0.747	0.497	0.095	0.031
8	45	-0.790	0.520	0.096	0.025
8	50	-0.879	0.417	0.086	0.029
8	60	-0.732	0.547	0.094	0.027

Table A8 (Continue)

Industry	# of Topics	Coherence Score			Stability Score
		u_mass	c_uci	c_npmi	Jaccard
9	30	-0.452	0.694	0.094	0.022
9	40	-0.572	0.618	0.089	0.024
9	45	-0.514	0.751	0.101	0.025
9	50	-0.606	0.659	0.092	0.030
9	60	-0.636	0.636	0.090	0.027
10	30	-0.321	0.761	0.101	0.038
10	40	-0.572	0.670	0.095	0.034
10	45	-0.631	0.644	0.092	0.030
10	50	-0.651	0.711	0.102	0.029
10	60	-0.678	0.678	0.095	0.029
11	30	-0.299	0.698	0.095	0.020
11	40	-0.350	0.614	0.087	0.020
11	45	-0.374	0.546	0.078	0.022
11	50	-0.472	0.640	0.093	0.019
11	60	-0.528	0.587	0.084	0.019
12	30	-0.434	0.835	0.105	0.016
12	40	-0.562	0.869	0.109	0.016
12	45	-0.513	0.925	0.111	0.015
12	50	-0.624	0.865	0.109	0.017
12	60	-0.649	0.794	0.106	0.016

For example, Figure A3 plots the topic coherence measures and the topic stability measure for each LDA model in industry 1 (Nondurables). The ideal number of topics was decided by a large value of  $c\_uci$  and a smaller value of Jaccard similarity. Therefore, the LDA model with 45 topics was the preferred model in this study. The same method was applied to other industries to choose the preferred model.



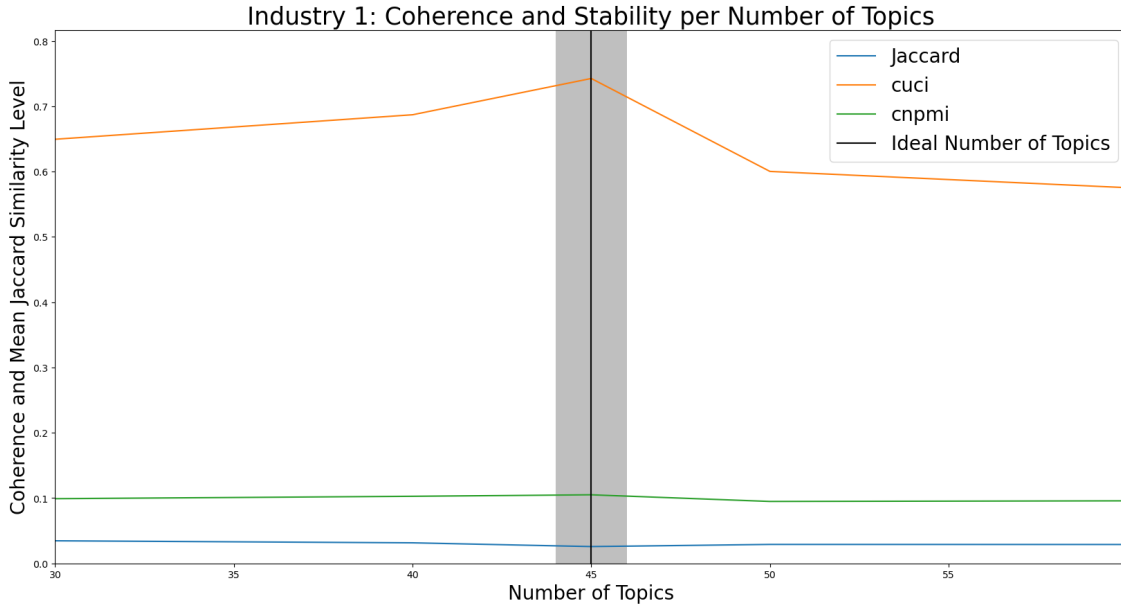


Figure A3 Selecting the Ideal Topic Numbers for Industry 1 (Nondurables)

In this study, I did not utilize the  $c_v$  measure because the calculation of this measure exceeds my available computing resources. Table A9 shows the time used to calculate the  $c_v$  measure for three industries. The machine used to calculate the measure was the Linux VM with 80 processors and 1024GB RAM. It took 156.5 hours to calculate the  $c_v$  measures for the six LDA models. The computing speed slows down as the number of topics increases.

Table A9 Information About Calculating the C\_V Coherence Score

Machine	# of Processors	RAM (GB)	Industry	Documents	# of Topics	C_V	Hours
Linux VM	80	1024	1	23,445	30	0.659	2
Linux VM	80	1024	1	23,445	40	0.642	8
Linux VM	80	1024	1	23,445	45	0.651	9
Linux VM	80	1024	1	23,445	50	0.626	12.5
Linux VM	80	1024	1	23,445	60	0.618	18
Linux VM	80	1024	2	12,241	30	0.613	1
Linux VM	80	1024	2	12,241	40	0.596	2
Linux VM	80	1024	2	12,241	45	0.609	4
Linux VM	80	1024	2	12,241	50	0.606	7
Linux VM	80	1024	2	12,241	60	0.591	8
Linux VM	80	1024	3	40,578	30	0.651	6.5
Linux VM	80	1024	3	40,578	40	0.655	13
Linux VM	80	1024	3	40,578	45	0.646	13
Linux VM	80	1024	3	40,578	50	0.617	24
Linux VM	80	1024	3	40,578	60	0.625	28.5
Total							156.5

#### 5.4 LDA Model Application

After selecting the preferred model for each industry, I applied the model to extract topics for each 10K/10Q document and earnings conference call transcript. Then, I kept the main topics based on the topic probability that is higher than 0.01. After determining the main topics for each document, I calculated the topic overlap for each pair of 10K/10Q filing and the conference call transcript.

Table A10 shows the top 10 topics for industry 1 (Nondurables). Figure A4 plots a 2D visualization of the 45 topics in industry 1. Figure A5 plots a 3D visualization of the 45 topics in industry 1. Appendix C lists all the topics for each industry.

Table A10 The Top 10 Topics for Industry 1 with 45 Ideal Topics

Topic #0	segment credit hallwood group energy facility industry manufacturing carpet revenue service book segment client education learn publishing acquisition
Topic #1	technology executive employment termination employee payment party benefit time
Topic #2	information day
Topic #3	apparel store brand retail license group wholesale klein inventory customer participant benefit employee payment account retirement service employer committee pension
Topic #4	vf brand diamond consumer price growth hershey benefit nut distribution
Topic #5	price bunge corn sugar seed commodity crop soybean plant sell
Topic #6	reddy paper ice certain manufacturing bahama capital industry unifi machine
Topic #7	china prc subsidiary use party ltd co exchange foreign rmb
Topic #8	wine brand vineyard february constellation grape hill file mcgraw spirit
Topic #9	juneregistrant six accounting disclosure three first sfas officer reporting
Topic #10	

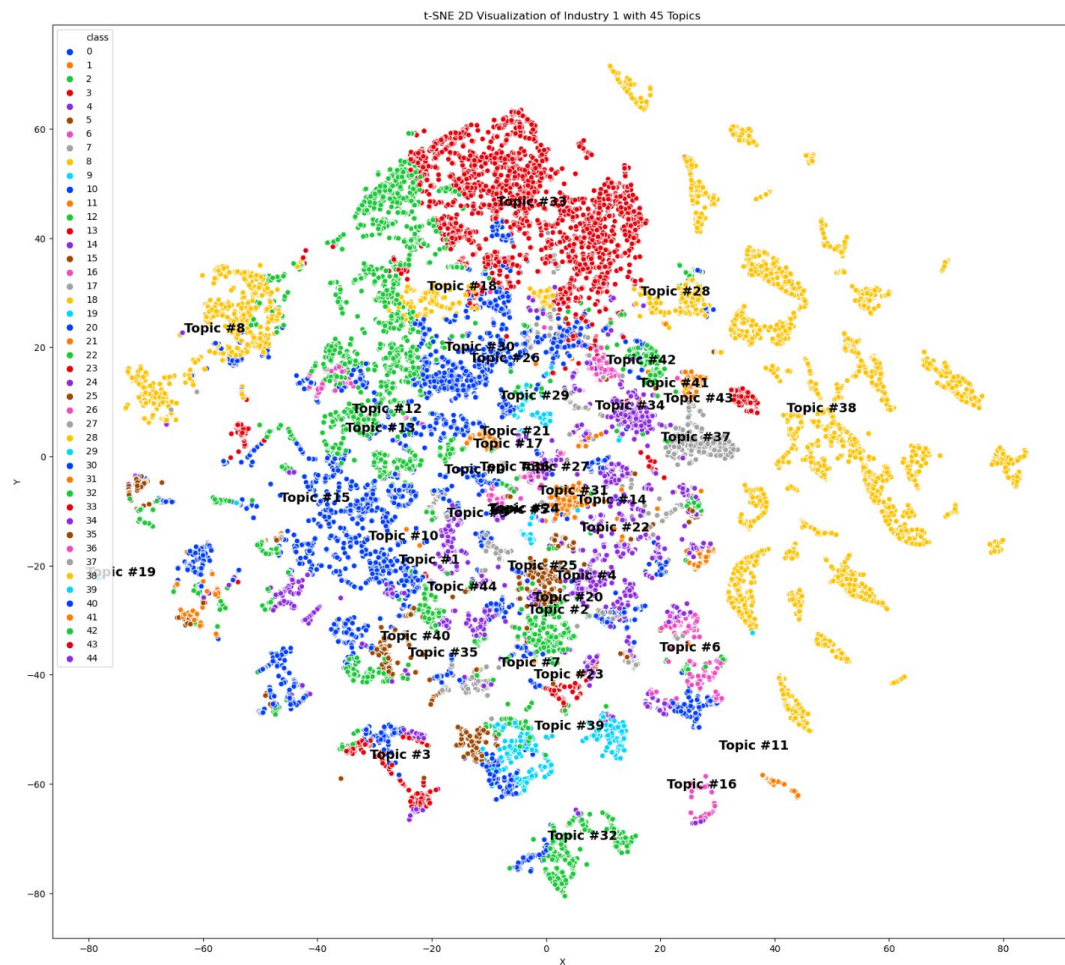


Figure A4 t-SNE 2D visualization of the 45 topics in industry 1 (Nondurables)

t-SNE 3D Visualization of Industry 1 with 45 Topics

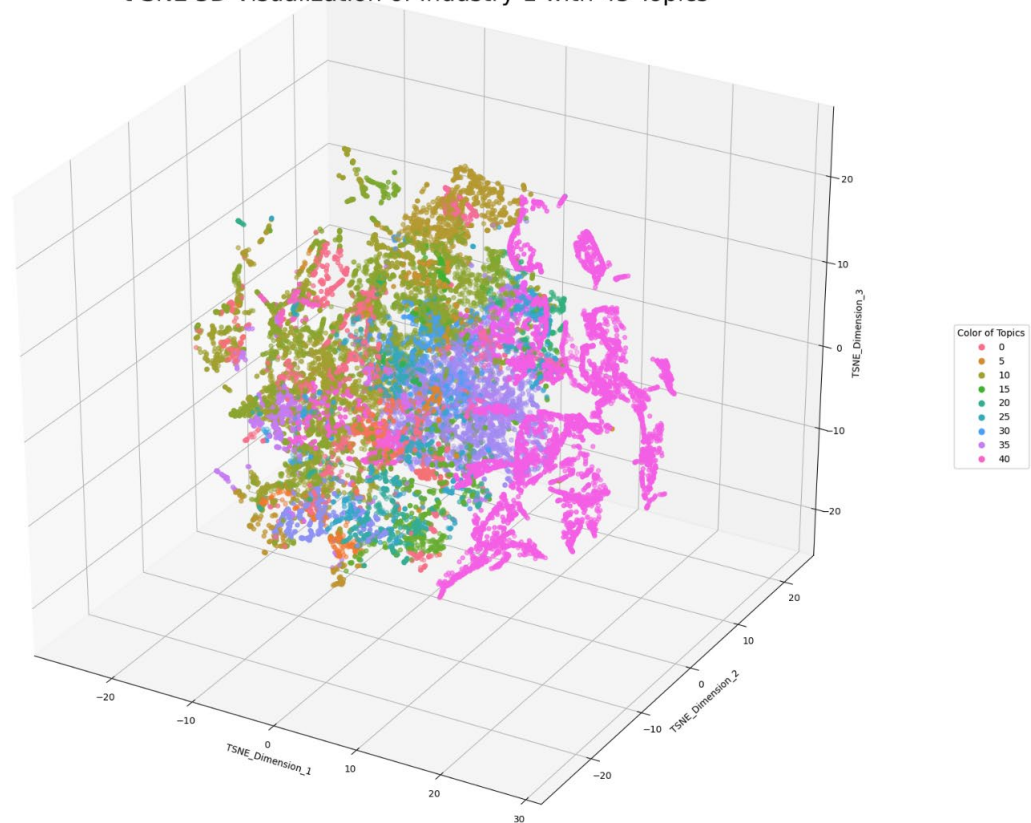


Figure A5 t-SNE 3D visualization of the 45 topics in industry 1 (Nondurables)

## 6. Comparison of LDA Training Time Between Tomotopy and Gensim

In this section, I compared the performance of the Tomotopy's LDA model and Gensim's LdaMulticore model based on three industries. Specifically, I trained 5 models with different numbers of topics for each industry by utilizing both methods. I specified same hyperparameters and iterations for both methods to enable the comparison. Table A11 reports the comparison results. Gensim took four times the total time to train the 15 LDA models than Tomotopy. Therefore, I suggest scholars to train the LDA models by utilizing the Tomotopy library if their corpus is not extremely large.

Table A11 Comparison of LDA Training Time Between Tomotopy and Gensim

Machine	Industry	Documents	# of Topics	Iterations	Tomotopy (Hours)	Gensim (Hours)
Widows VM	6	78,683	30	1000	1.5	3.3
Widows VM	6	78,683	40	1000	1.3	7.4
Widows VM	6	78,683	45	1000	1.6	8.4
Widows VM	6	78,683	50	1000	1.9	8.3
Widows VM	6	78,683	60	1000	1.7	9.6
Widows VM	9	47,339	30	1000	1.0	1.8
Widows VM	9	47,339	40	1000	0.9	4.1
Widows VM	9	47,339	45	1000	1.0	4.4
Widows VM	9	47,339	50	1000	1.2	4.5
Widows VM	9	47,339	60	1000	1.1	4.8
Widows VM	10	58,244	30	1000	1.5	2.7
Widows VM	10	58,244	40	1000	1.3	5.8
Widows VM	10	58,244	45	1000	1.7	6.7
Widows VM	10	58,244	50	1000	2.0	7.0
Widows VM	10	58,244	60	1000	1.8	7.3
Total					21.3	86.0

## 7. Conclusions

In this Appendix, I described the machines I utilized in this study, the data cleansing process, the document similarity calculation processes and computer runtime, and LDA application processes and computer runtime in each process. For scholars who are planning to apply the LDA method to financial reports or other text data in their projects, I give the following suggestions.

(1) Do the lemmatization instead of stemming when cleansing their data if they have enough time and computing power.

(2) If possible, utilize the cloud computing services (e.g., Amazon Web Services and OpenStack) to get access to a virtual machine with a high memory and more processors in their “big text data” analysis.

(3) Divide the sample to subsample based on industries or year range when applying the LDA method to financial reports.

(4) Utilize the Tomotopy library to apply the LDA model if their corpus is not extremely large, and they have access to machines with large memory.

In this study, I did not compare the computer runtime and output quality of different topic modeling methods. Comparing different topic modeling methods may help us choose the proper methods to extract meaningful topics from firm's disclosures such as SEC filings, corporate social responsibility reports, and earnings conference call transcripts. I plan to compare the performance of the LDA model with principal component analysis (PCA) and non-negative Matrix Factorization (NMF) in the future.

## Reference

- Ball, C., Hoberg, G., & Maksimovic, V. (2015). Disclosure, business change and earnings quality. Available at SSRN 2260371.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What Are You Saying? Using topic to Detect Financial Misreporting. *Journal of Accounting Research*, 58(1), 237-291. <https://doi.org/10.1111/1475-679x.12294>
- 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(2), 221-245. <https://doi.org/10.1016/j.jacceco.2017.07.002>
- Huang, A. H., Lehavy, R., Zang, A. Y., & Rong, Z. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach [Article]. *Management Science*, 64(6), 2833-2855. <https://doi.org/10.1287/mnsc.2017.2751>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
- 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*.

### APPENDIX 3. INDUSTRY TOPICS SUMMARY

I list all the topics for each industry here. The industry is based on Fama-French 12 industry classification.

#### Industry 1 with 45 Topics (Nondurables)

Number	Topics
Topic #1	segment credit hallwood group energy facility industry manufacturing carpet
Topic #2	revenue service book segment client education learn publishing acquisition technology
Topic #3	executive employment termination employee payment party benefit time information day
Topic #4	apparel store brand retail license group wholesale klein inventory customer
Topic #5	participant benefit employee payment account retirement service employer committee pension
Topic #6	vf brand diamond consumer price growth hershey benefit nut distribution
Topic #7	price bunge corn sugar seed commodity crop soybean plant sell
Topic #8	reddy paper ice certain manufacturing bahama capital industry unifi machine
Topic #9	china prc subsidiary use party ltd co exchange foreign rmb
Topic #10	wine brand vineyard february constellation grape hill file mcgraw spirit
Topic #11	june registrant six accounting disclosure three first sfas officer reporting
Topic #12	tobacco cigarette inc altria group usa court pm state subsidiary
Topic #13	registrant reporting march three disclosure act officer information item exchange
Topic #14	fruit fresh dole banana produce inc subsidiary game food de
Topic #15	director accounting inc compensation officer management item reporting estimate executive
Topic #16	service revenue cintas acquisition segment customer check operating due provider
Topic #17	cola coca bottle beverage percent bottler volume territory refer unit
Topic #18	beverage drink pepco energy brand kraft distributor distribution bottle volume
Topic #19	common inc development service energy issue management director subsidiary loss
Topic #20	september nine third october three fair november investment respectively revenue
Topic #21	store retail brand inventory wholesale footwear golf apparel consumer sell
Topic #22	mattel toy revenue game license brand entertainment international hasbro inventory
Topic #23	food brand frozen acquisition snack consumer customer sell price retail

Industry 1 (Continue)

Number	Topics
Topic #24	brand revenue licensee license royalty licensing retail trademark approximately lauren
Topic #25	store brand retail morris franchise rocky heinz carter factory decrease
Topic #26	award option grant performance restrict exercise common vest unit committee
Topic #27	technology patent health research revenue warrant animal use development license
Topic #28	could future customer condition price risk ability significant adversely affect
Topic #29	partnership land property water farm revenue fund real acre estate
Topic #30	brand beer brewing brewery craft coors molson distribution state sell
Topic #31	common warrant issue convertible price conversion per prefer purchase option
Topic #32	corporation director board meeting member stockholder class officer person time
Topic #33	price food feed beef egg facility production chicken protein approximately
Topic #34	officer common director exchange act accounting item management reporting inc
Topic #35	party seller buyer closing purchaser right respect set purchase law
Topic #36	credit facility loan senior certain subsidiary debt table acquisition content
Topic #37	coffee food dairy price milk brand dean facility green organic
Topic #38	lender loan borrower agent credit administrative party obligation time document
Topic #39	think go see growth well look get first last question
Topic #40	revenue advertising medium newspaper service new news publishing digital operating
Topic #41	food segment brand commodity bakery price benefit percent week foodservice
Topic #42	tenant lease landlord premise property lessee rent right building lessor
Topic #43	holder right trustee upon notice indenture act time pursuant law
Topic #44	court tobacco case plaintiff state defendant file district claim damage
Topic #45	foreign currency segment impact hedge fair operating loss benefit exchange



Industry 2 with 40 Topics (Durables)

Number	Topics
Topic #1	wabco fiscal robot government ceramic contract september technology development
Topic #2	september nine approximately compare segment facility third acquisition sell operating
Topic #3	delphi automotive system technology benefit vehicle certain liability content claim
Topic #4	fiscal revenue customer consumer audio technology electronics new march approximately
Topic #5	game revenue machine casino new table license lease fiscal system
Topic #6	customer could director reporting officer management inc item internal registrant
Topic #7	director corporation holder board meeting right time notice person stockholder
Topic #8	ford credit billion vehicle automotive loss service receivables truck high
Topic #9	vehicle facility fiscal cooper percent credit customer benefit commercial production
Topic #10	executive employee employment termination benefit payment party time provision law
Topic #11	lender agent loan borrower credit party obligation administrative time subsidiary
Topic #12	light energy lead fiscal revenue system customer technology service contract
Topic #13	fiscal furniture store retail design operating segment home retailer consumer
Topic #14	gm billion vehicle motor general subsidiary benefit certain gmac table
Topic #15	stanadyne engine corporation holding subsidiary navistar benefit truck segment pension
Topic #16	think go see well look first would get question margin
Topic #17	industry fiscal rv dealer vehicle unit home percent manufacture price
Topic #18	honeywell international sealy tempur mattress segment claim impact bedding unit
Topic #19	fair credit table segment foreign liability certain facility content loss
Topic #20	icahn investment mogul federal enterprise segment fund cvr railcar certain
Topic #21	whirlpool safety esw group facility affinia holding currency certain credit
Topic #22	week facility yankee candle senior store credit unit kei retail
Topic #23	fiscal display customer decrease revenue segment service brady light due
Topic #24	motorcycle harley davidson retail service hdfs global dealer finance inc
Topic #25	party seller lease tenant landlord buyer purchaser closing right property
Topic #26	vehicle facility program truck production dana customer wheel manufacturing benefit

Industry 2 (Continue)

Number	Topics
Topic #27	fiscal oil mining mine energy mineral approximately fuel gas project
Topic #28	participant award option committee grant benefit employee payment restrict performance
Topic #29	automotive facility corporation benefit visteon restructuring certain production content system
Topic #30	china prc subsidiary bank ltd co account loan foreign government
Topic #31	trust claim court guc new bankruptcy action gm debtor distribution
Topic #32	registrant three june reporting march act disclosure officer condense information
Topic #33	common warrant issue price convertible option per conversion prefer fair
Topic #34	officer common director accounting management item exchange act reporting disclosure
Topic #35	holding visant fiscal service senior facility certain corporation corp jostens
Topic #36	fiscal autoliv modine percent segment brand high water operating currency
Topic #37	vehicle fuel system electric model zap could automotive service technology
Topic #38	fiscal vehicle segment corporation order rockford equipment fire contract defense
Topic #39	sfas fiscal accounting credit facility option approximately subsidiary inc first
Topic #40	jewelry moissanite inventory customer jewel sell approximately consumer franchise finish

Industry 3 with 60 Topics (Manufacturing)

Number	Topics
Topic #1	customer technology revenue semiconductor equipment corn development process manufacturing
Topic #2	director officer mr common executive inc board management security item
Topic #3	fiscal acquisition industrial aerospace segment group operating corporation component system
Topic #4	novelis intevac price metal alcan aluminum bway inc march roll
Topic #5	paper packaging mill appleton price paperboard high benefit approximately fiber
Topic #6	lease tenant landlord lessee premise property lessor rent right notice
Topic #7	percent tool acquisition co segment industrial ltd electric system europe
Topic #8	copper cable wire price mueller sell general inventory metal purchase
Topic #9	ge investment billion service loss loan capital security segment receivables
Topic #10	steel price metal ton facility scrap mill nucor state roll
Topic #11	executive participant benefit employee payment employment termination service time account
Topic #12	foreign segment currency impact benefit hedge loss fair table derivative
Topic #13	debenture common conversion holder issue convertible price security day principal
Topic #14	facility senior subsidiary loan certain debt inc acquisition table revolve
Topic #15	common warrant price issue convertible prefer option per conversion purchase
Topic #16	party seller closing buyer respect right set law forth purchase
Topic #17	packaging food container plastic crown beverage plant customer holding resin
Topic #18	water system commercial technology service residential itt pump air new
Topic #19	contract government system program service defense atk snap fiscal corporation
Topic #20	vessel contract project revenue offshore facility mcdermott construction work marine
Topic #21	kodak digital image printing service patent technology information film revenue
Topic #22	march registrant reporting act three exchange officer information internal disclosure
Topic #23	china prc subsidiary account ltd exchange currency party foreign co
Topic #24	aircraft contract program boeing commercial service system revenue production government
Topic #25	project contract revenue service power segment facility gas energy nuclear
Topic #26	machine firearm new state inventory service international line device taser

Industry 3 (Continue)

Number	Topics
Topic #27	system technology segment service acquisition customer equipment industry management due
Topic #28	brand store retail consumer fiscal sell wholesale currency inventory distribution
Topic #29	greenbrier lease railcar graftech august table exchange content service subsidiary
Topic #30	agent lender loan borrower administrative party obligation time subsidiary respect
Topic #31	wind turbine revenue power fiscal energy development health contract march
Topic #32	think go see look get well would first question growth
Topic #33	service acquisition print customer printing solution graphic revenue technology charge
Topic #34	award option grant participant performance committee restrict right exercise unit
Topic #35	could customer future condition ability significant price require risk affect
Topic #36	power energy system solar technology fuel development revenue service project
Topic #37	construction concrete price cement plant operating segment owen corn approximately
Topic #38	oil service revenue gas drilling rig well activity equipment technology
Topic #39	boat dealer marine fiscal polaris industry percent brunswick retail new
Topic #40	venture joint ntic fiscal service foreign technology equity option compare
Topic #41	equipment percent engine dealer fiscal caterpillar due international price inventory
Topic #42	claim court asbestos settlement liability file bankruptcy case insurance action
Topic #43	paper verso newpage price fiber holding facility coat debt mill
Topic #44	tire goodyear rubber titan neenah mold technology clark kimberly raw
Topic #45	receivables purchaser account clause group party bank seller de relevant
Topic #46	revenue imax film system theater digital arrangement lease worldwide service
Topic #47	paper pulp mill canadian price canada dollar wood facility production
Topic #48	coal mining coke mine suncoke partnership production energy facility approximately
Topic #49	fair revenue liability fiscal estimate accounting reporting asu loss require
Topic #50	aluminum price facility contract zinc metal aleris power primary production
Topic #51	reporting inc item form registrant internal director officer table estimate
Topic #52	september june nine six registrant three compare condense decrease officer

Industry 3 (Continue)

Number	Topics
Topic #53	armstrong awi inc nmhg industry facility floor nacco truck subsidiary
Topic #54	building facility price tech october fiscal aluminum segment associate system
Topic #55	fiscal sfas accounting option fair issue fasb estimate require liability
Topic #56	corporation director board meeting holder time security person trustee officer
Topic #57	fiscal new segment furniture griffon facility percent trailer prior corporation
Topic #58	berry plastic corporation fiscal september group senior acquisition facility hold
Topic #59	alcoa metal alloy price high titanium aerospace raw nickel facility
Topic #60	railcar lease equipment crane unit revenue fleet service industry car

Industry 4 with 45 Topics (Energy)

Number	Topics
Topic #1	petroleum drilling license development option area work block field
Topic #2	plan employee executive participant award employment termination benefit payment time
Topic #3	director officer mr executive board compensation audit committee management service
Topic #4	pipeline gathering midstream partnership system facility unit partner segment energy
Topic #5	fiscal refining united product table crude content consolidated petroleum retail
Topic #6	exploration block contract eog development petroleum government area contractor crude
Topic #7	accounting sfas consolidated option fair liability issue change related reporting
Topic #8	bankruptcy plan lien facility energy certain chapter senior linn reorganization
Topic #9	partnership partner registrant general manage limited reef item accounting prove
Topic #10	quarter go think get see look question would first come
Topic #11	lender agent borrower loan administrative party time obligation document bank
Topic #12	corporation director board meeting member person time stockholder indemnitee right
Topic #13	amendment file reference exhibit inc form resource incorporate amend llc
Topic #14	could regulation future state business condition law affect material act
Topic #15	june six three registrant quarter compare condense decrease officer act
Topic #16	pbf refinery crude llc product energy inventory refining barrel unit
Topic #17	drilling lease work produce drill exploration development field prospect prove
Topic #18	common warrant convertible prefer issue per conversion series purchase option
Topic #19	holder indenture trustee registration subsidiary transfer act exchange right upon
Topic #20	product cvr facility crude business refining unit fertilizer approximately nitrogen
Topic #21	march registrant three act reporting officer exchange disclosure information internal
Topic #22	basin apache plan drilling north qep bakken form dakota williston
Topic #23	trust trustee royalty unit distribution underlie sandridge receive unitholders profit
Topic #24	corporation percent billion plan quarter crude table first sale benefit
Topic #25	refinery crude product refining fuel refine pipeline sale segment facility
Topic #26	court claim file plaintiff lawsuit district action defendant consolidated state

Industry 4 (Continue)

Number	Topics
Topic #27	partnership atlas mgp partner revenue derivative fair market future liability
Topic #28	prove future reservoir development drilling data produce quantity table operating
Topic #29	business common inc officer issue exploration exchange issuer management energy
Topic #30	september nine three registrant quarter condense compare decrease act officer
Topic #31	sale barnwell mineral development energy fiscal land revenue current payment
Topic #32	unit partner partnership general distribution common llc limited acquisition unitholders
Topic #33	coal mine mining ton sale table energy content contract customer
Topic #34	technology product business venture project development plant joint china fuel
Topic #35	contractor contract lng work tenant facility construction landlord project lease
Topic #36	fund manager project ridgewood energy related llc capital investment management
Topic #37	fair liability consolidated accounting loss reporting related item revenue change
Topic #38	rig drilling contract noble offshore rate operating subsidiary revenue market
Topic #39	partnership partner limited general ii geodyne sale energy manage registrant
Topic #40	texas common drilling ford eagle approximately exploration county shale energy
Topic #41	derivative facility rate table contract commodity hedge fair per borrowing
Topic #42	service revenue drilling rig customer segment equipment business contract facility
Topic #43	party seller buyer closing right respect title set purchase obligation
Topic #44	data seismic service revenue customer acquisition client facility equipment geophysical
Topic #45	offshore gulf mexico exploration vessel related table content anadarko facility

Industry 5 with 45 Topics (Chemicals)

Number	Topics
Topic #1	executive award employee benefit employment option grant payment termination
Topic #2	sfas option accounting fiscal inc account grant officer approximately registrant
Topic #3	director officer common mr security executive board management act accounting
Topic #4	technology project common waste development warrant energy issue per option
Topic #5	huntsman international facility subsidiary llc corporation loss chemical debt continued
Topic #6	director corporation holder board security right time meeting person common
Topic #7	flavor gamble procter fragrance celanese growth hercules care impact currency
Topic #8	facility senior holding certain subsidiary ebitda service table performance secure
Topic #9	option technology per inc grant june revenue common new director
Topic #10	facility production development amyris could total isobutanol certain patent fuel
Topic #11	revenue technology development could patent research license customer future develop
Topic #12	de license fragrance licensee inter parfums le licensor brand inc
Topic #13	fiscal scott unilever gro jdi miracle holding senior inc purchase
Topic #14	oil gas revenue well service technology apio landec production fiscal
Topic #15	phosphate potash mosaic mine corporation production mining inc fertilizer tonne
Topic #16	common warrant issue convertible conversion per purchase option exercise security
Topic #17	september june nine three six segment first compare table change
Topic #18	fiscal brand care consumer category new growth skin retail store
Topic #19	project owner construction contract work design coal plant contractor carbon
Topic #20	claim court grace settlement bankruptcy asbestos liability file case defendant
Topic #21	kronos tio certain nl facility benefit change subsidiary production future
Topic #22	polymer facility kraton chemical segment eastman performance revenue approximately operating
Topic #23	ppg chemical fiscal segment facility acquisition site environmental certain remediation
Topic #24	solutia monsanto seed percent fmc agricultural trait crop segment charge
Topic #25	think go see look growth well would first question get
Topic #26	cabot chemical facility dupont certain lsb work plant chemours el



Industry 5 (Continue)

Number	Topics
Topic #27	china prc fertilizer production account june ltd revenue subsidiary approximately
Topic #28	party right seller law notice respect write set obligation forth
Topic #29	ethanol corn plant grain energy approximately production distiller llc fiscal
Topic #30	facility chemical lyondell ethylene westlake millennium equistar production high table
Topic #31	polyone olin alkali facility chlor water environmental caustic pension soda
Topic #32	partnership partner facility nitrogen unit general gas fertilizer natural ammonia
Topic #33	lincolnway avon fiscal impact foreign due representative exchange currency new
Topic #34	dow percent corporation chemical asbestos insurance claim union liability carbide
Topic #35	agent trustee security payment respect indenture mean time clause document
Topic #36	technology option energy development battery inc system power application grant
Topic #37	revlon corporation inc credit senior facility loan subsidiary certain consumer
Topic #38	currency praxair impact gas charge program benefit foreign operating air
Topic #39	customer fair liability change reporting estimate table could loss future
Topic #40	lender borrower loan agent credit administrative party time obligation document
Topic #41	biodiesel fuel plant production diesel oil feedstock gallon facility reg
Topic #42	march registrant reporting three exchange act disclosure officer information internal
Topic #43	lease tenant property landlord right mortgage premise foot lessee say
Topic #44	segment specialty currency foreign high corporation chemical benefit facility additive
Topic #45	unit management nalco member service subsidiary purchase llc acquisition transfer

Industry 6 with 50 Topics (Business Equipment)

Number	Topics
Topic #1	data law awardee employer rsus la bank country en
Topic #2	inventory sell manufacturing order account component supplier warranty equipment credit
Topic #3	dealer vehicle automotive system consumer data navigation mobile gps marketing
Topic #4	borrower lender loan agent credit bank obligation document administrative collateral
Topic #5	healthcare health client system solution software medical care patient hospital
Topic #6	march condense nine six disclosure compare unaudited procedure chief decrease
Topic #7	system sensor digital segment equipment test application sell printer monitoring
Topic #8	storage data network system partner software channel server support solution
Topic #9	property oil exploration gas acquisition mineral investment mining acquire one
Topic #10	confidential supplier request exhibit contractor file order work write treatment
Topic #11	digital video client content box medium project cable television india
Topic #12	seller respect closing purchaser law buyer transfer transaction set forth
Topic #13	board compensation audit mr annual file incorporate grant reference exhibit
Topic #14	facility manufacturing electronics component segment design high environmental program restructuring
Topic #15	corporation board meeting stockholder person notice certificate vote class law
Topic #16	patent memory license flash royalty intel licensee drive venture micron
Topic #17	israel audio israeli video dollar communication device total development design
Topic #18	network wireless communication provider access solution mobile carrier voice telecommunication
Topic #19	semiconductor design wafer device test distributor high foundry manufacturing application
Topic #20	game title license software development platform entertainment release online royalty
Topic #21	internet subscriber com marketing online website web domain name consumer
Topic #22	brand spectrum insurance segment fgl acquisition holding subsidiary battery risk
Topic #23	satellite system contract network telesat lottery loral launch communication dish
Topic #24	software approximately digital solution image system development wave mobile application
Topic #25	instrument research system science scientific life acquisition development agilent laboratory
Topic #26	sfas approximately compensation grant record investment fasb effective disclosure recognize

Industry 6 (Continue)

Number	Topics
Topic #27	meter system smart utility ivoice class water memc qad project
Topic #28	employment termination benefit day provision law release write claim without
Topic #29	warrant convertible conversion prefer series per holder exercise upon principal
Topic #30	card transaction solution system ncr processing bank merchant software fee
Topic #31	software license recognize arrangement contract development defer fee support maintenance
Topic #32	user advertising content mobile search medium online consumer website advertiser
Topic #33	software license application support maintenance development solution total acquisition enterprise
Topic #34	court file claim patent action district complaint settlement litigation defendant
Topic #35	credit facility loan debt acquisition senior subsidiary capital revolve covenant
Topic #36	power solar energy project system battery contract cell development manufacturing
Topic #37	test development research medical patent clinical license fda device system
Topic #38	display laser application apply development system high manufacturing research semiconductor
Topic #39	think go see look well growth get question first would
Topic #40	foreign currency investment content impact acquisition segment benefit record primarily
Topic #41	nortel contract bankruptcy canadian debtor creditor claim court canada network
Topic #42	optical communication manufacturing high network rf component application design wireless
Topic #43	tenant landlord lease premise building rent day property notice lessee
Topic #44	disclosure procedure pursuant capital development file small principal rule issuer
Topic #45	contract government system program communication defense segment corporation award agency
Topic #46	participant award grant exercise unit committee restrict subject vest performance
Topic #47	percent billion hp currency benefit segment earnings software pension contract
Topic #48	china pre ltd subsidiary co currency loan limited law chinese
Topic #49	ability adversely affect subject addition harm law risk property additional
Topic #50	solution data subscription client platform cloud growth software offering application

Industry 7 with 60 Topics (Telecom)

Number	Topics
Topic #1	network carrier local telephone communication fcc telecommunication line long
Topic #2	director board officer mr corporation committee executive meeting compensation stockholder
Topic #3	court claim action file district certain settlement complaint class plaintiff
Topic #4	contract related receivables purchase receivable certain servicer funding acquisition payment
Topic #5	frontier deltacom corporation facility itc debt call communication credit telecommunication
Topic #6	call patent voip global provider product vonage number phone fiscal
Topic #7	alaska gci wireless segment network communication ac access facility member
Topic #8	product network technology solution software sale system syniverse support development
Topic #9	xm radio satellite sirius subscriber holding content cox music subscription
Topic #10	satellite intelsat network launch hughes hn new certain equipment system
Topic #11	usa mobility fox state software related united belo primarily message
Topic #12	quarter think go see growth look well get question first
Topic #13	nextel brazil network de mexico currency subscriber operating handset dollar
Topic #14	windstream debt benefit qwest network pension certain centurylink due access
Topic #15	sfas accounting option fair income estimate net loss compensation liability
Topic #16	network fiber level data communication carrier acquisition paetec telecommunication capital
Topic #17	sm npac provider subscription version data soa block number local
Topic #18	advertising medium program television channel russian tv license cme group
Topic #19	hotel interactive system room sale entertainment product game guest ntn
Topic #20	comcast cable warner twc program content network nbcuniversal income segment
Topic #21	executive employment employee termination payment benefit party provision right without
Topic #22	fair income estimate liability reporting related accounting net recognize loss
Topic #23	participant award committee benefit employee option grant payment mean determine
Topic #24	subscriber satellite echostar dish directv network certain program related patent
Topic #25	cable video system program franchise mediacom broadband llc partnership subscriber
Topic #26	china prc subsidiary exchange limited ltd currency income issue equity

Industry 7 (Continue)

Number	Topics
Topic #27	sale fiscal product international account approximately counsel telecommunication due related
Topic #28	sprint network subscriber wireless nextel clearwire pc lease device spectrum
Topic #29	could future ability significant condition material operating information affect additional
Topic #30	tower site lease wireless mobile operator sba communication related rental
Topic #31	cellular tds wireless partnership license could related certain operating income
Topic #32	wireless license fcc spectrum network communication leap cricket alltel carrier
Topic #33	cablevision program holding csc net network new certain contract msg
Topic #34	station television broadcast nexstar program fcc advertising local broadcasting mission
Topic #35	party contractor confidential request information exhibit work day right write
Topic #36	charter holding operating subsidiary llc debt cco credit capital cable
Topic #37	party seller closing buyer respect purchaser right material law transaction
Topic #38	tivo product subscription content january technology fiscal sale related development
Topic #39	wireless verizon network data billion benefit segment related primarily pension
Topic #40	network program content television entertainment cbs distribution medium film advertising
Topic #41	outdoor channel clear senior advertising subsidiary facility certain due credit
Topic #42	facility agent clause group finance bank document lender party relevant
Topic #43	radio station fcc advertising license broadcast broadcasting one operating medium
Topic #44	tenant landlord lease premise building rent property lessee center ibx
Topic #45	network telecom new access income carrier telephone table content equipment
Topic #46	station television radio program medium advertising broadcast fcc broadcasting license
Topic #47	june september three nine six registrant condense decrease unaudited quarter
Topic #48	fiscal idt property corporation telecom april october july energy straight
Topic #49	lin tv venture television joint station loan table medium shortfall
Topic #50	program hallmark channel medium crown subscriber advertising holding card playboy
Topic #51	issue warrant convertible price per conversion principal purchase option debenture
Topic #52	trustee indenture subsidiary holder issuer restrict payment guarantor person respect

Industry 7 (Continue)

Number	Topics
Topic #53	warrant holder exercise upon right registration price notice number conversion
Topic #54	prefer series dividend convertible holder issue price conversion outstanding right
Topic #55	liberty group starz broadband subsidiary upc ii medium certain series
Topic #56	virgin medium limited ntl senior cable uk breda network segment
Topic #57	credit facility table senior debt content loan certain due subsidiary
Topic #58	march registrant three officer act reporting exchange disclosure information material
Topic #59	lender loan borrower agent credit administrative party subsidiary obligation respect
Topic #60	satellite redact launch terrestar contract network product msv data system

Industry 8 with 60 Topics (Utilities)

Number	Topics
Topic #1	water plant bvi oglethorpe cooperative government contract oc purchase
Topic #2	participant benefit employee payment account employer committee retirement time election
Topic #3	project plant trust llc share geothermal inc investment solar wind
Topic #4	pipeline natural etp partner morgan kinder llc williams transportation unit
Topic #5	nisource utility indiana customer electric minnesota investment transmission columbia share
Topic #6	utility corporation sce pg edison cpuc california international customer regulatory
Topic #7	partner partnership unit general distribution common natural limited gathering midstream
Topic #8	entergy louisiana state system new gulf corporation nuclear arkansas fuel
Topic #9	water utility gswc share customer stock contract approximately state wastewater
Topic #10	pseg pse holding new contract electric investment public generation due
Topic #11	bond trustee indenture security series principal payment holder trust redemption
Topic #12	aep subsidiary plant management opco swepco due transmission risk apco
Topic #13	party transmission employee day time system schedule operating work capacity
Topic #14	electric edison teco con dte tampa utility new customer cecony
Topic #15	foot say line thence property county lessee north bond west
Topic #16	tva pse puget natural electric contract customer washington derivative risk
Topic #17	llc calpine calgen certificate center lp generate index plant limited
Topic #18	duke carolina progress inc ohio llc indiana florida file corporation
Topic #19	mge electric wisconsin central group natural hudson customer utility share
Topic #20	consumer cm centerpoint electric michigan natural business utility houston customer
Topic #21	pepco phi dpl ace customer holding electric electricity distribution due
Topic #22	month june september three reporting march quarter material information disclosure
Topic #23	pnm new pnmr mexico tnmp subsidiary resource unit texas nmprc
Topic #24	nu electric cl nstar transmission psnh contract new wmeco distribution
Topic #25	ameren illinois electric ue missouri ip genco cilco cips customer
Topic #26	washington laclede wgl utility customer natural fiscal holding inc september

Industry 8 (Continue)

Number	Topics	
Topic #27	sempra sdg natural utility california socialgas cpuc project contract commodity	
Topic #28	midamerican pacificcorp customer og state contract due nevada high natural	
Topic #29	natural customer storage utility contract margin share regulatory new operating	
Topic #30	fpl dynegy nee approximately group inc nep contract nextera loss	
Topic #31	sfas accounting approximately sale issue file purchase court march settlement	
Topic #32	dp contract dpl customer retail risk hedge business derivative loss	
Topic #33	executive award stock performance share termination employment payment employee compensation	
Topic #34	tep aps un electric west pinnacle arizona purchase table capital	
Topic #35	nsp utility npc sppc minnesota xcel psc electric purchase nevada	
Topic #36	quarter go think would look well see get question earnings	
Topic #37	corp oncor holding efh tceh txu texas debt subsidiary efih	
Topic #38	generation exelon come peco table combine contract bge electric customer	
98	Topic #39	cleco corporation lpsc customer fuel acadia due information louisiana unit
Topic #40	partner natural pipeline unit processing product ngl partnership operating volume	
Topic #41	epa rule emission state fuel environmental impact regulation court new	
Topic #42	firstenergy ohio generation transmission fe file jcp new met ed	
Topic #43	benefit fuel electric nuclear liability use plant unit fair operating	
Topic #44	production oil questar natural pipeline reserve well corporation share property	
Topic #45	mirant nrg genon generation sce america llc contract eme atlantic	
Topic #46	aep management subsidiary risk contract flow sale plant tcc cspco	
Topic #47	southern georgia mississippi alabama additional information matter fuel gulf psc	
Topic #48	utility ugi natural propane fiscal customer subsidiary sale electric unitil	
Topic #49	dominion idaho virginia idacorp ipc avista customer project pge generation	
Topic #50	director stock corporation share board meeting shareholder time person prefer	
Topic #51	could reporting officer accounting item material management information act liability	
Topic #52	agent collateral issuer party document payment security respect finance mean	



Industry 8 (Continue)

Number	Topics
Topic #53	contractor owner work party project test equipment construction site use
Topic #54	ppl supply electric lg ku contract subsidiary lke risk information
Topic #55	ipl electric alliant wpl utility hei heco hawaiian asb loan
Topic #56	lender borrower agent administrative loan bank time obligation issue respect
Topic #57	allegheny ae supply generation firstenergy certain fe pjw transmission virginia
Topic #58	kcp lng plain great cheniere liquefaction pas sabine kansa missouri
Topic #59	stock share common business system product customer sale development inc
Topic #60	party seller respect buyer closing purchaser obligation right forth set

Industry 9 with 45 Topics (Shops)

Number	Topics
Topic #1	inventory market december supply acquisition facility equipment industrial service
Topic #2	common warrant issue convertible price conversion per prefer option holder
Topic #3	distributor december sell market marketing health use new manufacturing approximately
Topic #4	fair liability estimate loss related impairment december use accounting record
Topic #5	loan receivables service servicer account collection related balance fee payment
Topic #6	partnership oil unit fuel partner gas general energy service facility
Topic #7	court claim action file plaintiff district state complaint settlement class
Topic #8	party seller respect purchaser right closing purchase law set obligation
Topic #9	service party confidential information supplier use bank license gap program
Topic #10	revenue service customer online website com marketing internet advertising december
Topic #11	trust cemetery funeral service revenue investment preneed contract merchandise market
Topic #12	could customer market future condition ability price operating affect risk
Topic #13	registrant reporting act exchange officer internal disclosure information material march
Topic #14	vehicle ashland part auction auto service acquisition facility sell automotive
Topic #15	lease tenant landlord premise property rent lessee right lessor day
Topic #16	lender agent loan borrower party administrative obligation respect document collateral
Topic #17	officer director common management development accounting exchange issue item act
Topic #18	club warehouse sysco inc factory burlington new membership coat innophos
Topic #19	executive employment employee termination payment benefit party day follow provision
Topic #20	registrant sfas accounting disclosure reporting exchange act material officer internal
Topic #21	vehicle use new dealership automotive service manufacturer retail unit finance
Topic #22	llc sonic inc name new la line drive limited amendment
Topic #23	september nine october third november compare condense decrease week table
Topic #24	airgas gas acquisition inc operating national table welder prior march
Topic #25	june six first table july decrease compare facility three content
Topic #26	senior facility subsidiary holding loan debt secure certain inc capital

Industry 9 (Continue)

Number	Topics
Topic #27	restaurant franchise food new lease franchisees operating menu december week
Topic #28	merchandise inventory retail january week lease new card sell comparable
Topic #29	director compensation december inc officer executive table option board mr
Topic #30	corporation director board meeting stockholder person officer shareholder notice vote
Topic #31	participant benefit contribution account employee employer service payment distribution committee
Topic #32	game entertainment water new revenue video medium hollywood project december
Topic #33	partner partnership ferrellgas propane general unit operating distribution price customer
Topic #34	coffee franchise revenue retail brand approximately franchisees operating new international
Topic #35	think go see look get well question last would first
Topic #36	china december prc ltd exchange subsidiary revenue hong kong currency
Topic #37	service customer revenue software vendor technology solution segment account acquisition
Topic #38	beauty salon jewelry sally diamond sell appliance tiffany supply shop
Topic #39	foreign currency international segment exchange impact dollar related hedge operating
Topic #40	pharmacy service drug health care medical pharmaceutical prescription revenue program
Topic #41	facility laundry customer december inventory sell apparel service boat acquisition
Topic #42	food market retail customer distribution benefit operating item tobacco approximately
Topic #43	brand customer new continue growth category program initiative focus marketing
Topic #44	award option grant participant restrict performance committee unit vest exercise
Topic #45	metal price steel december facility inventory customer usa acquisition sell

Industry 10 with 30 Topics (Healthcare)

Number	Topics
Topic #1	research license technology trial vaccine patent cancer grant manufacturing
Topic #2	tenant landlord lease premise rent building property day notice lessee
Topic #3	loan lender borrower agent interest credit obligation subsidiary document respect
Topic #4	candidate approval patent regulatory trial future obtain third fda additional
Topic #5	drug pfizer fda pharmaceutical research license treatment approval milestone patent
Topic #6	service contract health client care medical healthcare management provider program
Topic #7	device tissue implant medical fda technology spine surgical procedure wound
Topic #8	registrant three june september march reporting officer disclosure exchange act
Topic #9	pharmaceutical generic fda patent drug new price approval certain net
Topic #10	china customer tax income pharmaceutical account prc sell new asset
Topic #11	executive employee employment termination benefit follow day provision compensation change
Topic #12	service hospital facility patient care medicare health rate center interest
Topic #13	license research patent drug collaboration technology program therapeutic gene milestone
Topic #14	warrant price issue convertible prefer series purchase holder exercise conversion
Topic #15	device system medical technology heart patient procedure catheter fda abbott
Topic #16	director officer option inc accounting board management compensation item reporting
Topic #17	contract dialysis government amgen treatment manufacturing fda program patient approval
Topic #18	license collaboration research milestone trial drug candidate program royalty patent
Topic #19	patient trial study phase drug treatment data cancer fda disease
Topic #20	drug pharmaceutical fda trial research study patent technology license approval
Topic #21	system device customer medical technology new sell procedure fda service
Topic #22	blood baxter plasma royalty license manufacturing approval fda certain united
Topic #23	eye sanofi treatment approval state receive patient manufacturing fda marketing
Topic #24	option award participant corporation director grant board exercise committee law
Topic #25	quarter think go look question see first get well growth
Topic #26	tax asset rate income consolidated net acquisition credit fair certain
Topic #27	court file claim district action complaint settlement litigation state plaintiff
Topic #28	license respect information confidential write pursuant material applicable set mean
Topic #29	test diagnostic laboratory service customer diagnostics technology image cancer new
Topic #30	facility lease property llc community living nursing senior care mortgage

Industry 11 with 30 Topics (Money)

Number	Topics
Topic #1	borrower section party agent seller respect obligation time document
Topic #2	fund equity fee revenue client related group certain compensation transaction
Topic #3	lease operating tenant real estate debt approximately cost venture unit
Topic #4	prefer holder section series trustee corporation payment class dividend right
Topic #5	real estate reit distribution fee advisor lease manager offering stockholder
Topic #6	health care member plan medical contract provider healthcare state provide
Topic #7	land sale price gas cost royalty oil production project development
Topic #8	derivative hedge mortgage table collateral agency swap obligation portfolio instrument
Topic #9	revenue customer product technology option inc common cost use sale
Topic #10	could operation condition future reporting subject act ability affect state
Topic #11	corporation form exhibit file item registrant act reference incorporate exchange
Topic #12	trading fund series future contract class manage commodity position advisor
Topic #13	insurance life benefit product annuity policy contract account derivative liability
Topic #14	plan executive section participant employee benefit award employment termination provide
Topic #15	fund contract future index commodity price exchange fee trading manage
Topic #16	trust mortgage trustee series certificate pool standard servicer minimum account
Topic #17	receivables finance portfolio billion consumer card related lease balance sale
Topic #18	september june nine six three quarter decrease compare table due
Topic #19	partnership partner limited general unit operating fund local sale distribution
Topic #20	quarter think go see look would well growth get question
Topic #21	mortgage sale related residential home purchase repurchase balance borrower held
Topic #22	criterion transaction applicable mortgage servicer compliance item respect pool platform
Topic #23	deposit federal total allowance institution real estate mortgage portfolio follow
Topic #24	common director combination officer acquisition warrant stockholder public mr issue
Topic #25	court claim file action plaintiff complaint settlement state certain district
Topic #26	march three registrant reporting disclosure information operation exchange use material
Topic #27	deposit total table portfolio allowance quarter first commercial compare balance
Topic #28	insurance premium reinsurance claim reserve write policy estimate ratio subsidiary
Topic #29	hotel operating revenue sale approximately resort operation llc cost debt
Topic #30	registrant reporting officer operation accounting disclosure internal act material exchange

Industry 12 with 45 Topics (Other)

Number	Topics
Topic #1	participant employee award employment termination benefit option payment grant
Topic #2	credit facility march table debt senior three content loan net
Topic #3	product sale material corporation aggregate project sand production state price
Topic #4	product development research clinical patent license drug trial technology sale
Topic #5	hotel property resort room franchise sale fee operating brand own
Topic #6	marketing advertising product medium online customer website internet content network
Topic #7	inc holding llc subsidiary merger certain file senior acquisition group
Topic #8	membership member fitness senior inc credit account mr executive accounting
Topic #9	quarter think go see look get well growth would question
Topic #10	waste facility environmental landfill disposal inc liability site recycle closure
Topic #11	sale equipment product rental customer store inventory sell new retail
Topic #12	fuel transportation freight customer operating logistics equipment per carrier driver
Topic #13	trustee indenture trust issuer holder global guarantor payment subsidiary principal
Topic #14	student program education school institution university title iv state enrollment
Topic #15	tenant lease landlord premise rent lessee property building lessor right
Topic #16	vessel charter ship cruise shipping marine day facility fleet contract
Topic #17	china prc subsidiary ltd exchange limited co foreign group currency
Topic #18	exploration property mineral claim mining gold option resource project stage
Topic #19	director mr executive board audit item act accounting exchange inc
Topic #20	corporation director board meeting stockholder holder person certificate right class
Topic #21	film entertainment production event theatre distribution music television fiscal park
Topic #22	option table compensation approximately net estimate fair quarter fiscal grant
Topic #23	home land sale mortgage loan community market real development estate
Topic #24	client health medical healthcare staff contract care insurance consult employee

Industry 12 (Continue)

Number	Topics
Topic #25	series class vehicle group ii hertz respect account rental car
Topic #26	currency foreign international global dollar exchange impact net segment benefit
Topic #27	energy power project technology plant generation solar price sale fuel
Topic #28	contract project construction work government estimate fiscal backlog engineering award
Topic #29	warrant issue price convertible conversion option per purchase holder exercise
Topic #30	mine mining gold production price per project ore sale silver
Topic #31	fair liability reporting accounting estimate loss asc asu disclosure entity
Topic #32	transaction payment card merchant processing travel fee customer consumer bank
Topic #33	rail railroad fuel bnsf coal railway claim due liability table
Topic #34	state facility per contract kforce inc own bingo ltd center
Topic #35	sfas accounting exchange disclosure act material inc registrant reporting information
Topic #36	june six three court quarter file registrant condense decrease claim
Topic #37	customer product technology software system solution sale data client new
Topic #38	aircraft airline fuel lease air airway flight engine united boeing
Topic #39	party right seller law respect notice write set buyer obligation
Topic #40	lender borrower loan agent credit administrative party obligation document respect
Topic #41	could market future ability condition risk significant change subject affect
Topic #42	september registrant nine three act reporting exchange disclosure information internal
Topic #43	game casino property facility nevada license vega entertainment la resort
Topic #44	oil gas unit partner pipeline partnership general energy distribution product
Topic #45	lease partnership equipment partner fund investment llc limited general manager

## REFERENCES

- Allee, K. D., & Deangelis, M. D. (2015). The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion. *Journal of Accounting Research*, 53(2), 241-274. <https://doi.org/10.1111/1475-679X.12072>
- Arif, S., Marshall, N. T., Schroeder, J. H., & Yohn, T. L. (2019). A growing disparity in earnings disclosure mechanisms: The rise of concurrently released earnings announcements and 10-Ks. *Journal of Accounting and Economics*, 68(1), 101-221. <https://doi.org/10.1016/j.jacceco.2018.11.002>
- Barker, R., Hendry, J., Roberts, J., & Sanderson, P. (2012). Can company-fund manager meetings convey informational benefits? Exploring the rationalisation of equity investment decision making by UK fund managers. *Accounting, Organizations and Society*, 37(4), 207-222. <https://doi.org/10.1016/j.aos.2012.02.004>
- Bowen, R. M., Davis, A. K., & Matsumoto, D. A. (2002). Do Conference Calls Affect Analysts' Forecasts? [Article]. *The Accounting Review*, 77(2), 285-316. <https://doi.org/10.2308/accr.2002.77.2.285>
- Brochet, F., Loumiotis, M., & Serafeim, G. (2015). Speaking of the short-term: disclosure horizon and managerial myopia. *Review of Accounting Studies*, 20(3), 1122-1163. <https://doi.org/10.1007/s11142-015-9329-8>
- Brochet, F., Naranjo, P., & Gwen, Y. (2016). The Capital Market Consequences of Language Barriers in the Conference Calls of Non-U.S. Firms [Article]. *The Accounting Review*, 91(4), 1023-1049. <https://doi.org/10.2308/accr-51387>
- Brown, L. D., Call, A. C., Clement, M. B., & Sharp, N. Y. (2015). Inside the “Black Box” of Sell-Side Financial Analysts. *Journal of Accounting Research*, 53(1), 1-47. <https://doi.org/10.1111/1475-679X.12067>
- Brown, L. D., Call, A. C., Clement, M. B., & Sharp, N. Y. (2019). Managing the narrative: Investor relations officers and corporate disclosure. *Journal of Accounting and Economics*, 67(1), 58-79. <https://doi.org/10.1016/j.jacceco.2018.08.014>
- Brown, N. C., Crowley, R. M., & Elliott, W. B. (2020). What Are You Saying? Using topic to Detect Financial Misreporting. *Journal of Accounting Research*, 58(1), 237-291. <https://doi.org/10.1111/1475-679x.12294>
- Brown, S., Hillegeist, S. A., & Lo, K. (2004). Conference calls and information asymmetry. *Journal of Accounting and Economics*, 37(3), 343-366. <https://doi.org/10.1016/j.jacceco.2004.02.001>
- Brown, S., Lo, K., & Lys, T. (1999). Use of R2 in accounting research: measuring changes in value relevance over the last four decades. *Journal of Accounting and Economics*, 28(2), 83-115. [https://doi.org/10.1016/S0165-4101\(99\)00023-3](https://doi.org/10.1016/S0165-4101(99)00023-3)
- Brown, S. V., & Tucker, J. W. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research*, 49(2), 309-346. <https://doi.org/10.1111/j.1475-679X.2010.00396.x>



- Burgoon, J., Mayew, W. J., Giboney, J. S., Elkins, A. C., Moffitt, K., Dorn, B., Byrd, M., & Spitzley, L. (2015). Which spoken language markers identify deception in high-stakes settings? Evidence from earnings conference calls. *Journal of Language and Social Psychology*, 35(2), 123-157. <https://doi.org/10.1177/0261927X15586792>
- Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic Complexity in Firm Disclosures: Obfuscation or Information? *Journal of Accounting Research*, 56(1), 85-121. <https://doi.org/10.1111/1475-679X.12179>
- Bushee, B. J., Matsumoto, D. A., & Miller, G. S. (2004). Managerial and Investor Responses to Disclosure Regulation: The Case of Reg FD and Conference Calls [Article]. *The Accounting Review*, 79(3), 617-643. <https://doi.org/10.2308/accr.2004.79.3.617>
- Calomiris, C. W., & Mamaysky, H. (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics*, 133(2), 299-336. <https://doi.org/10.1016/j.jfineco.2018.11.009>
- Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy Prices. *Journal of Finance*, 75(3), 1371-1415. <https://doi.org/10.1111/jofi.12885>
- Davis, A., Ge, W., Matsumoto, D., & Zhang, J. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2), 639-673. <https://doi.org/10.1007/s11142-014-9309-4>
- Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2015). The Determinants and Consequences of Information Acquisition via EDGAR. *Contemporary Accounting Research*, 32(3), 1128-1161. <https://doi.org/10.1111/1911-3846.12119>
- Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2016). The Usefulness of Historical Accounting Reports. *Journal of Accounting and Economics*, 61(2-3), 448-464. <https://doi.org/10.1016/j.jacceco.2015.12.001>
- Driskill, M., Kirk, M. P., & Tucker, J. W. (2020). Concurrent Earnings Announcements and Analysts' Information Production [Article]. *The Accounting Review*, 95(1), 165-189. <https://doi.org/10.2308/accr-52489>
- 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(2), 221-245. <https://doi.org/10.1016/j.jacceco.2017.07.002>
- Dzielinski, M., Wagner, A. F., & Zeckhauser, R. J. (2016). In no (un) certain terms: Managerial style in communicating earnings news. *SSRN Electronic Journal*.
- Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing [Article]. *Journal of Financial & Quantitative Analysis*, 52(2), 811-836. <https://doi.org/10.1017/S0022109017000187>

- Falkinger, J. (2008). Limited Attention as a Scarce Resource in Information - Rich Economies. *The Economic Journal*, 118(532), 1596-1620. <https://doi.org/10.1111/j.1468-0297.2008.02182.x>
- Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 15(4), 915-953. <https://doi.org/10.1007/s11142-009-9111-x>
- Gomez, E., Heflin, F., Lee, J. A., & Wang, J. (2018). Who and What Drive the Investor Response to Earnings Conference Calls? *SSRN Electronic Journal*.
- Huang, A. H., Lehavy, R., Zang, A. Y., & Rong, Z. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach [Article]. *Management Science*, 64(6), 2833-2855. <https://doi.org/10.1287/mnsc.2017.2751>
- Kimbrough, M. D. (2005). The Effect of Conference Calls on Analyst and Market Underreaction to Earnings Announcements [Article]. *The Accounting Review*, 80(1), 189-219. <https://doi.org/10.2308/accr.2005.80.1.189>
- Lee, J. (2016). Can Investors Detect Managers' Lack of Spontaneity? Adherence to Predetermined Scripts during Earnings Conference Calls [Article]. *The Accounting Review*, 91(1), 229-250. <https://doi.org/10.2308/accr-51135>
- Lee, Y.-J. (2012). The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices. *Contemporary Accounting Research*, 29(4), 1137-1170. <https://doi.org/10.1111/j.1911-3846.2011.01152.x>
- Lev, B., & Gu, F. (2016). *The End of Accounting and the Path Forward for Investors and Managers*. John Wiley & Sons.
- Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research*, 48(5), 1049-1102. <https://doi.org/10.1111/j.1475-679X.2010.00382.x>
- Louis, H., & Sun, A. (2010). Investor Inattention and the Market Reaction to Merger Announcements [Article]. *Management Science*, 56(10), 1781-1793. <https://doi.org/10.1287/mnsc.1100.1212>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
- Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What Makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion Sessions. *The Accounting Review*, 86(4), 1383-1414. <https://doi.org/10.2308/accr-10034>
- Mayew, W. J., Sethuraman, M., & Venkatachalam, M. (2020). Individual Analysts' Stock Recommendations, Earnings Forecasts, and the Informativeness of Conference Call Question and Answer Sessions. *The Accounting Review*, 95(6), 311-337. <https://doi.org/10.2308/tar-2017-0226>
- 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*,

Skinner, D. J. (2003). Should firms disclose everything to everybody? A discussion of “Open vs. closed conference calls: the determinants and effects of broadening access to disclosure”. *Journal of Accounting and Economics*, 34(1–3), 181-187. [https://doi.org/10.1016/S0165-4101\(02\)00074-5](https://doi.org/10.1016/S0165-4101(02)00074-5)

## VITA

### CHUANCAI ZHANG

#### EDUCATION

- Ph.D. in Business Administration-Accounting, Xiamen University
- Mater of Business Administration-Accounting, Dongbei University of Finance and Economics
- Bachelor of Business Administration-Accounting, Shandong Agricultural University

#### PROFESSIONAL POSITIONS HELD

- Research Assistant, University of Kentucky (2016-2021)
- Research Assistant, Hong Kong Baptist University (Oct 2014-Jun 2016)
- Research Assistant, City University of Hong Kong (Apr 2014-Aug 2014)

#### PROFESSIONAL PUBLICATIONS

- Zhang, M. C., D. Stone, and H. Xie. 2019. "Text Data Sources in Archival Accounting Research: Insights and Strategies for Accounting Systems' Scholars" *Journal of Information Systems* 33 (1):145-180.
- Zhang, C., and H. Chen. 2016. "Product market competition, state ownership and internal control quality" *China Journal of Accounting Studies* 4 (4):406-432.
- Zhang, C., and H. Chen. 2014. "Internal Control, Investor Sentiment and Stock Market Response to Earnings News" *China Economic Studies* 4: 61-74. In Chinese.
- Yang, D., C. Zhang, and H. Chen. 2014. "Internal Control, Integration Ability and M&A Performance: Empirical Evidence from Chinese Listed Firms" *Auditing Research* 3: 43-50. In Chinese.
- Chi, G., C. Zhang, and H. Han. 2012. "Individual Investors' Perception of Risk Associated Internal Control Deficiencies Disclosure: An Experimental Study" *Auditing Research* 2: 105-112. In Chinese.