A Textual Analysis of Financial Disclosure; Evidence from the Stock Exchange of Thailand

Mr. Siriyos Chuthanondha

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Economics) in Economics
Common Course
FACULTY OF ECONOMICS
Chulalongkorn University
Academic Year 2019

การประยุกต์ใช้ Textual Analysis กับรายงานผลการดำเนินงานของบริษัทจดทะเบียนใน
ตลาดหลักทรัพย์แห่งประเทศไทย

นายศิริยศ จุฑานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาเศรษฐศาสตรดุษฎีบัณฑิต
สาขาวิชาเศรษฐศาสตร์ ไม่สังกัดภาควิชา/เทียบเท่า
คณะเศรษฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Thesis Title	A Textual Analysis of Financial Disclosure; Evidence from the Stock Exchange of Thailand
By	Mr. Siriyos Chuthanondha
Field of Study	Economics
Thesis Advisor	Assistant Professor PONGSAK LUANGARAM, Ph.D.

Accepted by the FACULTY OF ECONOMICS, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy (Economics)

.... Dean of the FACULTY OF ECONOMICS

(Associate Professor SITTIDAJ PONGKIJVORASIN, Ph.D.)

DISSERTATION COMMITTEE

.... Chairman

(KAIPICHIT RUENGSRICHAIYA, Ph.D.)

.... Thesis Advisor

(Assistant Professor PONGSAK LUANGARAM, Ph.D.)

.... Examiner

(Associate Professor TANAPONG POTIPITI, Ph.D.)

.... Examiner

(Associate Professor SOTHITORN MALLIKAMAS, Ph.D.)

.... External Examiner

(Associate Professor Yuthana Sethapramote, Ph.D.)

ศิริยศ จุฑานนท์ : การประยุกต์ใช้ Textual Analysis กับรายงานผลการดำเนินงานของบริษัทจดทะเบียนในตลาดหลักทรัพย์แห่งประเทศไทย. (A Textual Analysis of Financial Disclosure; Evidence from the Stock Exchange of Thailand) อ.ที่ปรึกษาหลัก : ผศ. ดร.พงศ์ศักดิ์ เหลืองอร่าม

งานวิจัยนี้พยายามศึกษาหาความสัมพันธ์ระหว่างข้อมูลที่เป็นคำอธิบายในรูปของตัวอักษรกับผลตอบแทนของหลักทรัพย์โดยการใช้กระบวนการ Textual Analysis กับกลุ่มตัวอย่างในรายงานการเงินของบริษัทจดทะเบียนในตลาดหลักทรัพย์แห่งประเทศไทย ซึ่งฝ่ายจัดการของบริษัทนิยมจัดทำคำอธิบายและการวิเคราะห์ของฝ่ายจัดการ (Management Discussion and Analysis หรือ MD&A) เพื่อสื่อสารกับผู้ลงทุนและผู้มีส่วนได้ส่วนเสียของบริษัท โดยงานวิจัยนี้ได้นำวิธีการวิเคราะห์คำอธิบายในรูปของตัวอักษรมาสร้างเป็นดัชนีชี้วัดที่มีค่าเป็นตัวเลขที่สามารถนำค่าดัชนีมาเปรียบเทียบกันระหว่างบริษัทและเปรียบเทียบกันในแต่ละช่วงเวลาได้ จากการศึกษาพบว่าโดยเฉลี่ยแล้วนั้นฝ่ายจัดการเขียนรายงาน MD&A ครอบคลุมถึง 4 หัวข้อหลักๆ ได้แก่ 1) ผลการดำเนินงานของบริษัท 2) สถานะทางการเงินของบริษัท 3) ปัจจัยภายนอกบริษัทที่มีผลกระทบกับผลการดำเนินงาน และ 4) ปัจจัยภายในอุตสาหกรรมซึ่งบริษัททำธุรกิจอยู่ โดยฝ่ายจัดการจะเขียนถึงหัวข้อผลการดำเนินงานในเชิงบวกมากขึ้นหากคาดว่าผลการดำเนินงานของบริษัทในอนาคตจะเป็นไปในทิศทางสดใส อีกทั้งฝ่ายจัดการจะกล่าวถึงสถานะทางการเงินในอนาคตมากขึ้นหากคิดว่าบริษัทกำลังจะมีกำไรในทิศทางขาขึ้น อย่างไรก็ดีจากการศึกษาพบว่าผู้ลงทุนกลับให้ความสนใจกับทัศนคติของฝ่ายจัดการในคำอธิบายภายใต้หัวข้อสถานะทางการเงินของบริษัทและปัจจัยภายนอกที่ส่งผลกระทบกับกำไรบริษัทมากกว่าคำอธิบายในหัวข้อผลการดำเนินงานของบริษัท นอกจากนี้ผู้ลงทุนยังตอบสนองต่อคำที่มีความหมายเชิงบวกและเชิงลบไม่เท่ากัน โดยผู้ลงทุนให้ความสำคัญกับคำที่มีความหมายเชิงลบมากกว่า สังเกตได้จากผลตอบแทนของหลักทรัพย์ที่ลดลงมากเมื่อฝ่ายจัดการประกาศข่าวร้ายในคำอธิบายผลการดำเนินงาน อีกทั้งพบว่าคำที่มีความหมายเชิงบวกและเชิงลบนั้นยังส่งผลต่อความผันผวนของระดับราคาหลักทรัพย์ที่แตกต่างกัน สุดท้ายนี้หากผู้ลงทุนนำข้อมูลที่เป็นคำอธิบายในรูปของตัวอักษรมาใช้ประกอบการตัดสินใจในการลงทุนจะสามารถสร้างผลตอบแทนได้สูงกว่าอัตราผลตอบแทนเฉลี่ยของตลาด

สาขาวิชา	เศรษฐศาสตร์	ลายมือชื่อนิสิต	....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก	....

5985908129 : MAJOR ECONOMICS
KEYWORD: textual sentiment analysis, topic modeling, financial disclosures, future firms performance, abnormal stock returns, stock price volatility

Siriyos Chuthanondha : A Textual Analysis of Financial Disclosure; Evidence from the Stock Exchange of Thailand. Advisor: Asst. Prof. PONGSAK LUANGARAM, Ph.D.

This paper empirically explores causal relation between non-financial information and stock price performance by applying textual analysis to listed firms' financial disclosure in Thailand. Managers generally use language in Management Discussion & Analysis (MD&A) report to communicate value-relevant information to investors and other stakeholders. I applied various kinds of approach to quantify qualitative information. I found that, on average, management discussion reports four main topics, which are financial performance, financial status, external factor and industry specific topic. The result shows that managements discuss more proportion on financial performance topic with more positive net tone, when future ROA is increasing. Additionally, they tend to use ambiguous language in financial status topic, when managements expect firms' profitability to show an upward trend. More importantly, the result shows that investors place greater value on management tone in other topics rather than tone in financial performance. Particularly, this study revealed that investors reacted to this kind of information asymmetrically. The effect of the unfavorable tone of financial disclosures on stock market price was more pronounced and had more predictive power than the favorable tone. Finally, I found that net tones in MD&A report contain the critical information in predicting the stock return volatility.

Field of Study:	Economics	Student's Signature ....
Academic Year:	2019	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I would like to acknowledge my indebtedness and render my warmest thanks to my academic advisor, Dr. Pongsak Luangaram, who made this work possible. His kindly guidance and expert advice have been invaluable throughout all stages of the work. I would also wish to express my gratitude to Dr. Kaipichit Ruengsrichaiya and Dr. Yuthana Sethapramote for extended discussions and valuable suggestions which have contributed greatly to the improvement of the thesis. The thesis has also benefited from comments and suggestions made by Dr. Tanapong Potipiti and Dr. Sothitorn Mallikamas. I take this opportunity to thank them.

I am grateful The Stock Exchange of Thailand for the funding sources that allowed me to pursue my doctoral degree. I am forever indebted to the president of The Stock Exchange of Thailand, Dr. Pakorn Peetathawatchai, who has been very supportive throughout. His keen mind and abilities put me on the right track and enabled me to complete this research.

This thesis has been written during my work at the Research Department of The Stock Exchange of Thailand. This work would never have been completed without my supervisor. I would like to thank Dr. Soraphol Tulayasathien for providing excellent working conditions. Every result described in this thesis was accomplished with the help and support of fellow researchers and colleagues. On top of that, I would like to thank Mrs. Jiaranai Tantivess for reading and editing my final proof.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, who always pay attention to the education of their child. Without the fundamental of life they give me, I couldn’t come this far. Special thanks are due to my wife for her continuous support and understanding, and also giving birth of my first son in few coming months. My thanks are extended to my friends and administration office in Economics Faculty, Chulalongkorn University for helping me during these four wonderful years.

Siriyos Chuthanondha

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1
Introduction

The Efficient Market Hypothesis (EMH) states that securities prices were built by rational investors who correctly use all available information in evaluating their value, where the costs of establishing and processing information are explicitly known . However, the stock market is more complicated than that, and the stock price frequently fluctuates from its fundamental value. Therefore, there are many studies that try to answer which kind of information has the predictive power to explain stock returns. Earnings announcement is one the most important information materials that impact stock price. However, investors normally hurdle into the quantitative information like firms' revenue, net profit and cash flow, while qualitative information (usually in text format), which accounts for most of the total information, is left over. Many studies seemed to conclude that the weak or semi-strong form of inefficiency has recently vanished in developed markets. However, some researchers find that several developing markets are still semi-strong form of inefficiency. Many researches have tried to figure out whether the stock market can be forecast. According to the EMH, since news in nature happens randomly and is unknowable in the present, stock prices that combine all available information should follow a random walk pattern and the best bet for the next price is the recent price. However, many researchers continue to figure out about new information that can impact the firm's performance and the change in stock price. A growing body of literature empirically explores causal relations between qualitative data (text) e.g. annual report, news, and internet posting, etc. and stock market movements.

There are various kinds of document relevant to listed firms in the stock market. However, some of them contain information, and some of them do not. Despite security price reaction to information, it is not expected to be instantaneously completed. When the costs of obtaining and exploiting information are significant, investors who invest in resources in processing information may be compensated for the cost they incur and risks they bear. Their three main sources are financial reports written by managements, financial news and analyst forecast. If analyst and key items in financial report are inadequate or biased, measures of firms' textual variables may have additional predictive power for firms' future profit and returns.

The data volume around the world was surging dramatically, since our documents can be collected in the form of digitized. Perhaps, this immense volume is beyond human cognition, and poor understanding could result in bad decision making. We need new computational tools to help analyze, arrange, and apprehend this information, especially in text format. One of the earliest and the simplest methodologies in analyzing text is natural language processing (NLP). NLP is a tool to identify the sentence structure, grammar and part of speech. By using NLP, researcher can program the computer to categorize and tag specific words that they are interested and count those words from the entire document (as known as "Bag of words" approach). By using a "Bag of words" approach, the computer will know what type of word to look for, we can find a top frequent word, count target words or phrases and start to analyze as a term document matrix. From that point, technological progress allows researchers to parsing tons of documents with less time and less cost. We can command our personal computer to gauge a tone with dictionary base approach, search for main document theme by applying topic modeling and measure readability score of selected documents.

The Stock Exchange of Thailand (SET) index is one of the top indices with top performance in Asia for the past decade with highest liquidity, but SET index is also highly volatile. One explanation is that our Thai stock market is inefficient: investors have asymmetry information and many of them trade stock based on sentiment rather than fundamental. Therefore, this study will give a clearer picture with several contributions to the management communication through financial disclosure literature. First, to my knowledge, this paper will be the first that introduces textual analysis to parsing the large sample of quarterly data in financial disclosure of Thai listed company. Second, my study is one of the firsts in emerging stock market to use a Loughran and McDonald (LM) dictionary, a well-known dictionary for finance from the US, to count positive and negative words in English-Thai Management and Discussion Analysis (MD&A) documents. Finally, I extended the literature about whether managers communicate any information to investors by demonstrating that managers use explanation in terms of words throughout a financial report to signal their prospects for future firm performance, and that the investors respond to this signal.

Motivation

The speed of integrating information to stock valuation process depends on what type of information that investors use. There are at least two types of information usually found in earning announcement reports. One is “quantitative information” such as number, which is easier to collect and faster for investors to digest. Another is “qualitative information” which is text data that explains about each number representation. Since soft information is a lot more difficult to read and analyze, investors will need more time to interpret this kind of information. Although there are several attempts to analyze qualitative data in behavioral finance during the past decades, these papers are limited in small sample size . However, after global financial crisis in 2008, financial firms tend to downsize their organization to survive, and adopt the cutting edge technology to remain their competitiveness. Moreover, there are many professional institutes forecast that the volume of data in the world will be double in every two years from 2010 – 2020. With this speed of increase, human cannot understand all of the flow of data, and poor understanding lead to poor strategy. Fortunately, computational power of the computer also improved significantly, human can program machine to do such thing that people in the past cannot imagine. As a result, the word parsing technology has been developed dramatically that make researchers can analyze plenty of documents, which are kept in a digital format, with less time and less cost. As a result, extracting textual information from a large-sample of data may contain additional value to investors

Objectives

The advance in technology allows us to use machine to read and analyze this overwhelming data. Therefore, I would study about what information is inside financial disclosures released form listed companies in The Stock exchange of Thailand by using various kinds of textual analysis approaches. First, I would apply dictionary-based approach to measure textual sentiment in Management and Discussion Analysis (MD&A) reports that are published quarterly in English language. Then, I improve my model by applying complicated statistical model called Latent Dirichlet Allocation (LDA) to identify key themes hidden within the financial disclosure documents. Lastly, I employed EGARCH (1,1) model, one of the foremost techniques for modelling volatility in financial markets to test whether textual sentiment impact stock return volatility, with individual SET50 stock data during Jan 2012 – Mar 2019.

Significant

This issue involves with financial economics. Since in the stock market, information is very crucial asset, and one bad news can impact stock market and make market index goes down dramatically. Stock Exchange of Thailand (SET) is one of the top index performances in Asia for the past decade with highest liquidity, but SET index is also highly volatile. One explanation is that Thai stock market is inefficient, investors have asymmetry information and many of them trade stock based on sentiment rather than fact. Therefore, it is interesting to study how investors response to textual sentiment from firms' financial disclosure. Moreover, since qualitative data is magnificent and harder to digest by investors, to understand what inside the tons of words in financial document can be benefit for investors to form investment strategies and probably gain abnormal return. Policy makers could use textual analysis to measure the effectiveness of new rules and regulations. Furthermore, researchers can conduct studies with new qualitative information sources and other content analysis approaches that have not yet been widely used.

Contributions

This paper will be the first that introduces textual analysis to parsing the large data set in financial disclosure of Thai listed company. Although these textual analysis approaches have been proved that work well among developed stock markets, implementing them in emerging markets, where managements don't use English as primary language, different culture and writing style, might not guarantee the same result. For technical issue, my paper will be among the first in applying dictionary-based approach and topic modeling together in order to gauging the tone within the document topics. Last, I extend literature about management communication, and how Thai stock market response to this signal. Then we can gain benefit from understanding how listed companies communicate their performance to investors via financial disclosure reports, and what types of topic that they frequently mention.

The remaining parts of the paper are organized as follows: Chapter 2 reviews the literature about development of textual analysis in finance. Chapter 3 explains the samples of financial disclosures collection, textual analysis measurement and variable definitions. Chapter 4 constructs hypothesis and discusses methodology. Chapter 5 presents result of textual analysis and economics test results. Chapter 6 summarize and conclusions.

Chapter 2
Impact of textual sentiment to future firms' performance and stock markets response

Firm performance

Financial statement reports information relating the firm's performance, for example, current sales, direct and indirect cost and earnings which the companies aim to provide information to investors and shareholders during the period. Since they often use past managerial narrative disclosure to help assess the projections of a company. Therefore, managements should disclose information accountably in order to create trust among interested parties. However, there are some literatures, explaining about asymmetry information problems, which causes the market inefficiency and reduce trust among investors. For example, agency theory arises when there is asymmetry of information between the manager (agent) and the owner (principal). The manager has more information related to the company than the owner, but the manager often has difficulty in revealing their weaknesses. Another issue known as signaling theory which explains the behavior of managers that tend to express information only when firm shows favorable result. However, for managers of poorly performing companies, it would be difficult to disclose such unfavorable performance. To improve the creditability of disclosure, firm need to engage in good corporate governance.

Textual sentiment was proof that contains explanatory power to a future firm performance which is usually indicated by firm's profitability ratio e.g. return on asset (ROA) and return on equity (ROE). Davis, Piger et al. (2006) used return on asset (ROA) in multivariate regression to test the hypothesis that net optimistic language in earnings press releases is positively associated with future firm performance. They found evidence that managers used optimistic and pessimistic language in earnings press releases to provide investors with information about expected future firm performance. Li (2010) found that when managers are more optimistic when discussing future events in MD&A, future earnings and liquidity are indeed much better, even after controlling for stock returns and other predictors of future performance. Huang, Teoh et al. (2014) found that abnormal positive tone in the earnings press release is associated with poor future earnings and operating cash flows in each of one-year to three-year forward periods . Furthermore, Ferris, Hao et al. (2013) documented that prospectus conservatism for non-technology IPOs contains useful information about the firm's future operating performance.

Stock return predictability

Theory of efficient market hypothesis (EMH) introduced that the market is not anticipated. Therefore investors in stock market should not hope to gain a return that is significant higher than the market average. However, many studies document that most of the stock markets are still the weak or semi-strong form of inefficiency especially in emerging markets.

Therefore, researchers apply the event study to test the hypothesis that there is some information that has predictive power to stock returns in certain period of time while the events are normally company news, earning announcements and analyst report etc. This methodology has defined the dependent variable as the cumulative abnormal return (CAR) (or event period excess return) over some event window. Researchers can examine the tone of particular document by counting words that have positive and negative sentiment within the document. By doing so, researchers need to identify dictionary that will be used to create a word lists. There are couples of dictionaries that are popular among pioneer in this field including Harvard IV-4 and Diction. However, two of them are general English language linguistic dictionaries rather than dictionaries that are specific to the domain of financial document. Henry (2008), and Loughran and McDonald (2011) mentioned that tagging pessimistic words following by general English dictionaries might not be appropriate, because these words did not usually express negative meaning in financial terms. They did suggest their own list of specific words for the domain of financial disclosure (FD and LM negative word list respectively). Jegadeesh and Wu (2013) applying LM negative word list and find that tone of a policy announcement on the releasing day of the Federal Open Market Committee (FOMC) minute meeting is correlated with significantly higher stock market volatilities. Although, Loughran and McDonald negative word list becomes a predominant in textual analysis for finance, some literatures combine several dictionaries for examining tone of financial document in order to compare a predictive power of market response. For example, Davis, Ge et al.(2015) use three different wordlists i.e. DICTION, Henry (2008), and Loughran and McDonald (2011) wordlist to examine the effect of managerial “Style” on the tone of earnings conference calls.

Stock price volatilities

In common, there are many ways in which stock market volatility and market efficiency are linked. This makes volatility an important phenomenon to study. The practitioners and academics in finance have focused on modeling of stock market volatility since it can be used in forecasting the stock future returns. In addition, these models can be used in risk management, portfolio construction, derivative pricing etc. The most common volatility patterns studied include seasonal effects and volatility clustering.

The most common volatility measurement is standard deviation, however volatility clustering pattern violated the random walk assumption. Therefore, recent studies introduce new volatility measurements such as Auto Regressive Conditional Heteroskedasticiity (ARCH) class model, which implied that volatility today is the sum of square of volatility yesterday and the day before. Then GARCH model has been improved by adding long term variance, which represent mean reversion effect or unconditional variance.

There is a wide range of research that has studied numerous factors that may have an influence on volatility. This may include financial disclosure, earning press release and macroeconomic statements. The volatility-based studies pay attention to the transparency or accuracy of the disclosed information. Many literatures stated that the announcement effect of good and bad news are not equal, while negative tones seem to lower stock returns. However, the relationship between textual sentiment and stock volatilities could not find a robust result yet. Likewise, Das and Chen (2007) use textual analysis to measure sentiment in message board postings for 24 high-tech stocks. They find that stock message board postings are related to stock market levels, trading volume, and volatility. Moreover, in the previous studies, researchers proposed that it is better to distinguish between good and bad news by using an asymmetric model, such as quadratic or exponential Generalized Autoregressive Conditionally Heteroskedastic (GARCH) model.

Chapter 3
Data and Sample Selection

There are at least three types of qualitative data sources about earning announcement; firms' financial disclosure, media express (e.g. financial news) and analyst reports. I chose corporate disclosures (e.g. annual report, earning press release, conference call transcript and IPO filling) as primary sample. Since these documents were written by management team of the company, who has more inside information than investors. Firms' financial disclosures normally provide information related to firms' performance that is useful for firms' stakeholders such as investors, debtors and other creditors in order to make decision about allocating their fund to the company. Therefore, good managements would use financial disclosures to communicate valued information to earn trust from stakeholders.

Moreover, unlike numeric data that contain only past firm's performance, Securities and Exchange Commission of Thailand (SEC) encourage listed company to include forward looking about firm's operation in the financial report. Furthermore, corporate financial disclosure has well structural format, formal writing style, and certain period of announcement. Therefore, this source of data is frequently used to examine the textual sentiment from management perspective. The following literatures study about announcement effects on asset prices. Tetlock (2008), Engelberg (2008), Davis and Tama-Sweet (2011), Demers and Vega (2011), Durnev and Mangen (2009) concludes that the negative tone of corporate disclosures or changes in the negative tone from the recent past are significantly correlated with short window contemporaneous returns around the date that the disclosures are made. Only study of Jegadeesh and Wu (2013) indicate that there are significant relationship between their measure of the tone of 10-Ks and market reaction for both negative and positive words . Loughran, McDonald et al. (2009) find that the percentages of negative words in the S-1 are much more powerful variables in explaining levels of underpricing than many 4 commonly used IPO control variables. However, there are some drawbacks in corporate disclosure, since manager may not speak the whole truth and the corporate disclosures release in quarterly or annually basis. Furthermore, the corporate disclosure is quite long, focusing on specific part e.g. MD&A or forward looking statement might be adequate.

The Stock Exchange of Thailand (SET) was established in 1975 The SET is regulated by the SEC. It lists around 600 companies with a market capitalization of $569 billion in 2019. Significantly, Thai listed companies have improved in their quality and increasingly gained global recognition. There were 40 Thai listed companies added to MSCI Standard Index, a leading global index, while 20 Thai listed companies added to Down Jones Sustainability Index (DJSI), the highest number in ASEAN for 5 consecutive years. One factor that makes Thai listed firms to be selected these global indices is the quality of the information disclosure to stakeholders. SET has founded special team to encourage listed firms to disclose information complied with sustainability guideline continuously.

There are at least 3 types of corporate financial disclosures in SET database such as annual report, form 56-1 and management discussion and analysis (MD&A). After looking for entire 670 companies in SET and mai, listed before June 30, 2017, I found that 130 companies (accounted for 51% of total market capitalization) submitted English version of their annual report continuously from 2011 to 2018. Moreover, Thai listed companies usually published their MD&A instead of quarterly earnings press release, 49 companies (accounted for 50% of total market capitalization) release their MD&A every quarter from Q4/2011 – Q4/2018. Although the number of companies that have released MD&A continuously was less than companies that released annual report, MD&A releases were on a quarterly basis and a total of 29 earnings quarters were published, making the sample size of 1,421 samples. MD&A documents are the largest samples for Thai financial disclosure documents.

Table 1 List of firms published MD&A documents in English

Sector	#comp	Listed company symbol	#document
Agriculture	3	**MINT**, SAUCE, STA	84
Technology	5	**ADVANC, INTUCH, TRUE**, THCOM, SYMC, CSL	140
Resource	10	**PTTEP, EGCO, TOP, IRPC, RATCH, PTT, BCP**, LANNA, BAFS, ESSO	280
Finance	7	TCAP**, KKP, SCB, TMB, KBANK, KTB**, MFC, LHBANK	196
Service	13	**BH, AOT, BDMS,** BJC, **BTS**, PSL, MCOT, GENCO, **THAI**, ERW, TTA, GRAMMY, MODERN	364
Industry	4	**IVL, PTTGC**, VNT, LHK	112
Real Estate	5	**SCC, CPN**, MBK, DRT, TTCL	140

Note: the stock tickers which are highlighted in bold indicate that they are in SET50 index

The additional advantage is that it is noteworthy to study about information contained in them. Firstly, MD&A is the primary source of linguistic information for journalists and stock analysts to track Thai firms' quarterly performance. The listed firms would initially send their financial statement summary along with MD&A documents to report their financial performance to SET. Most of journalists and stock analysts have access to these materials within the same day after the firms have submitted. The first point is what managers explained in MD&A could be considered the primary source of listed companies' news and analyst reports, which have been spread to practically all investors in the short period of time. The second point is that SET has given the guideline to listed firms in order to submit MD&A documents, involving some major changes in the firm's profitability, and the fact that the Securities and Exchanges Commission (SEC) has encouraged managements to explain their view about the forward looking of future firm's performance. Therefore, it can be said that we can expect to discover their signal about firms' management outlook in these MD&A documents.

To study about impact of textual sentiment to future firm performance and stock market return in Thailand, I have divided the data into three groups as follows: 1) textual data, 2) accounting data and 3) financial market data. Most of the data are from SET database from 2012 to 2018.

Textual data

One of the earliest and the simplest methodologies in analyzing text is natural language processing (NLP). NLP is a tool to identify the sentence structure, grammar and part of speech. By using NLP, researcher can program the computer to categorize and tag specific words that they are interested and count those words from the entire document (as known as “Bag of words” approach). In the past, researchers might need to deal with many lines of codes manually in complicate programs like Perl and General Inquirer (GI), but recently they can use open-source programs like R and Python to examine millions of financial documents for free. By using a “Bag of words” approach, the computer will know what type of word to look for. With the main assumption that word is independence and a sequent of word doesn’t matter, we can find a top frequent word, counting target words or phrases and start to analyze as a term document matrix.

MD&A reports **Term-document matrix**

	ADVANC 1q2012	PTT 1q2012	KBANK 1q2012	SCC 1q2012
Profit	10	15	8	20
Expense	6	5	5	9
Income	20	25	24	10
Revenue	8	12	16	5
Cost	10	10	20	18
Sale	9	8	0	30

Words **Vectors**

Figure 1 Transforms word to number by constructing term-document matrix

When investigating these 49 companies that have been reporting MD&A every quarter during Q4/2011 – Q4/2018, it was found out that 24 companies were in SET-50 Index (as of Dec 31, 2018), which accounted for 44.5% of total market capitalization, while other 25 companies were not in the SET-50, including companies in mai. It is significant in terms of diversity. These 49 companies were from various industries, where resource and finance industries contain the most samples, accounting for over half of each industry's market capitalization. From 1,421 documents, firstly, I started the preprocessed text data by converting all documents from PDF files to text files, using PyPDF2 package in Python. There are total 6 million words in my samples. After numerical and punctuation were eliminated, my sample size consists of 4 million words. Next, I randomly read some documents and decided to remove words that were below 2 characters since all of them are meaningless in terms of financial perspective. Then I started programing the computer to count the most frequent words that appear on my sample. I found that words that related to currency (baht, THB, USD), unit (thousand, million, billion) and time (monthly, quarterly, annually) are in the top 30 most frequent words, but these kinds of words don't express anything relevant to sentiment. Therefore, I removed these three word categories and I also remove the stop words, a commonly used word (such as "article", "pronoun", "proposition") that should be programed to ignore, both when indexing entries for searching and when retrieving them as a result of a search query. My final sample contained 2,984,369 words.

After finishing preprocessing data, my final round sample had 2,114 words per document on average. The resource and finance sectors contained more words than average, because most of them were large-cap stocks, which having wider and deeper business, and many of them have more proportion of foreign holding than small-cap stocks (Figure 1). Moreover, the number of words has been quite stable over time from 2012 – 2018, and as I observed a slight increase in number of words in the fourth quarter each year (Figure 2), as managers would normally summarize the highlight of the year at year-end in the fourth quarter.

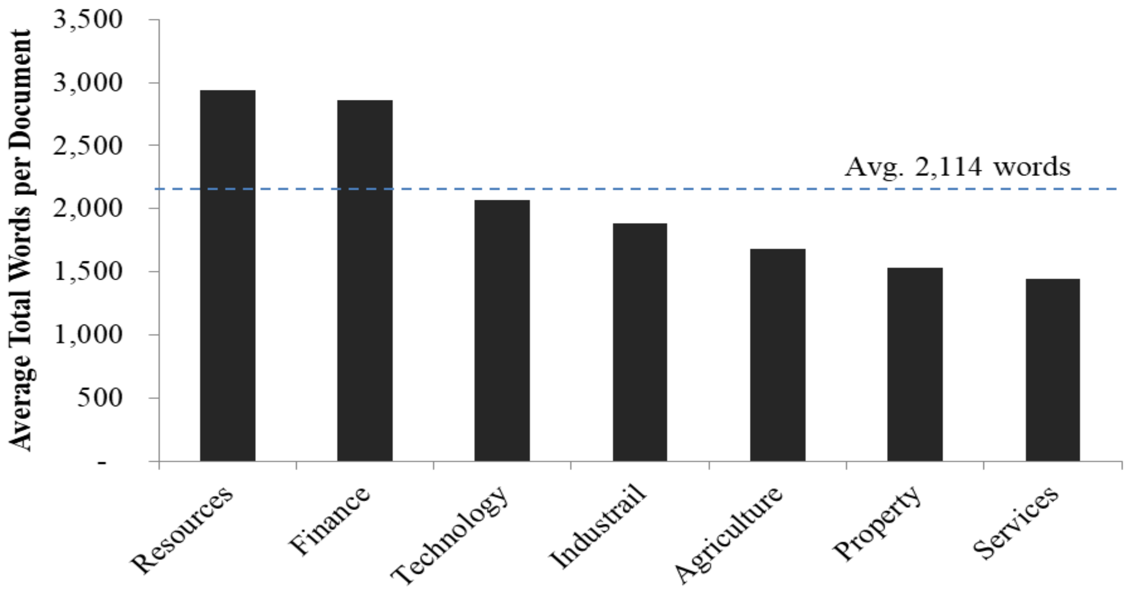

Figure 2 Average of total word count by industry

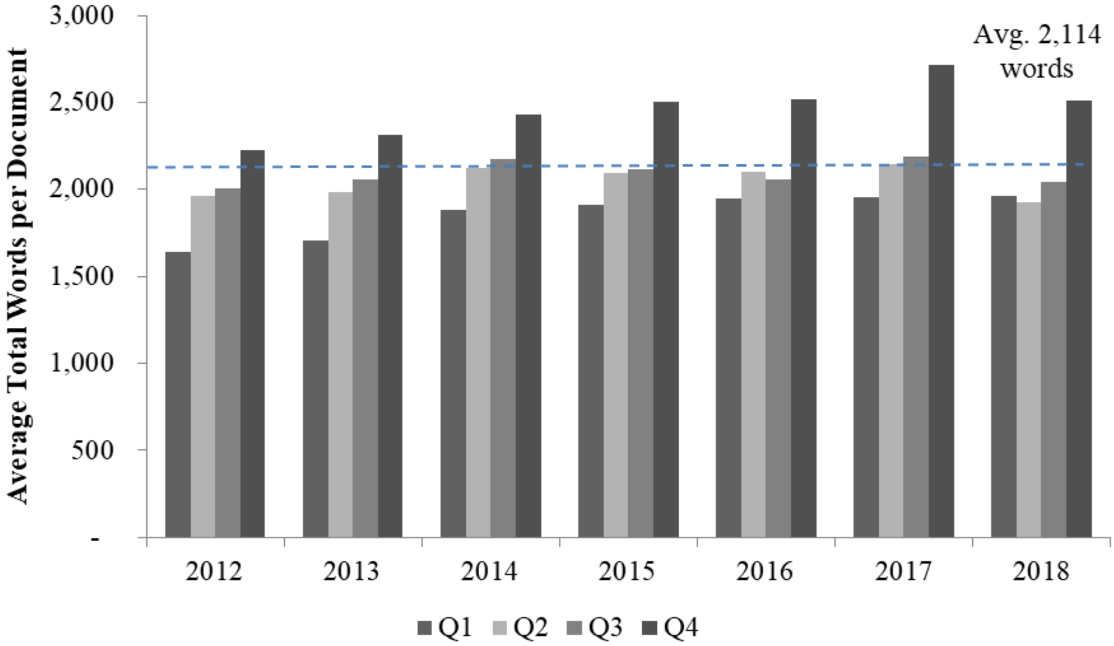

Figure 3 Average of total word count by year (breakdown to quarter)

After analyzing top 30 most frequent words in final sample, I found that the most common words in MD&A report are business term such as “analysis”, “company”, “discussion”, “public”, and “management”. Furthermore, the item in financial statement like “income”, “profit”, “sales”, “revenue”, “cash”, “loans”, “EBITDA”, “loss”, and “cost” were used primarily in these documents. It was also found that tonal words, which express about sentiment, such as “increase”, “decrease”, “higher”, and “growth” are among the top common words in MD&A documents (Figure 3). Since managers would report firm’s performance comparatively on the basis of quarter-on-quarter (QoQ) and year-on-year (YoY) basis, I expected to see the impact of textual sentiment changes in these documents.

Figure 4 Top 200 most frequent words from MD&A documents

Accounting data

For data that came from financial statement, I constructed a number of variables from SET and Bloomberg database. My dependent variables for my tests is future firm performance, which represents by the average of return on asset (ROA) in four following quarters after current financial disclosures were announced. The ROA is calculated as firms' net profit divided by total assets as of the end of each quarter. For control variables in my regressions, I collected variables possibly correlated to the future firm performance and the stock returns around the date of the MD&A report announcement, in order to measure for quantifiable information in financial disclosures, which couldn't capture by textual data alone. I used recent quarter sales (REV) and used its natural logarithm (LOGREV) as a measure of firm size. In my regression, I included ROA in order to capture persistence in firms' financial performance. I also included standard deviation of ROA in four following quarters after current financial disclosures were announced to control for the uncertainty in future firm performance. I also use controlled dummy variables such as 1) LOSS, whether the firm reported negative earning in specific quarters 2) DET_FS, whether managements wrote an explanation of firms' balance sheet and cash flow statement position in MD&A report each quarter 3) DIV_INC, whether firm has changed dividend payment during the event window 4) NONREC_POS whether firm had extraordinary profit in reported quarter and 5) NONREC_NEG whether firm had extraordinary loss in reported quarter. In addition, I added some standard firm's characteristics, which may affect future firm's performance differently, such as profit margin (PM) ratio, asset turnover (AT) ratio, debt to asset ratio (DA), and book to market ratio (BM).

Financial data

To measure market returns, I define the cumulative abnormal return (CAR) over today, the past "m" days and the next "n" days after the MD&A report release date CAR[-m, +n]. In conducting event study test, I calculated the abnormal return using market model where abnormal return equal to actual stock return minus required return (market return multiply by 360-day beta of individual stock). Since the sum of 3, 5, 7 days around announcement date depicts as CAR[-1,+1], CAR[-2,+2], CAR[-3,+3] respectively. If CAR = 0, indicates that there is no event that impacts stock return. In addition, I constructed a number of return variables to represent the level of market response such as sum of abnormal daily return for today as CAR[0,0], one day forwards (+1 days) as CAR[0,+1], three day (+3 days) as CAR[0,+3] and five day (+5 days) as CAR[0,+5] after announcement date from Bloomberg database.

Moreover, I controlled for the quantitative earning surprise by measuring the current quarter earnings surprise (SURP) as the difference between actual net profit and the most recent forecasted analyst consensus of net profit before earnings announcement, normalized by the most recent forecasted analyst consensus. I retrieve actual earning and consensus analyst earnings forecast from Bloomberg data base. Furthermore, I define the dummy variable BEAT to be 1 if announced net profit in the current quarter exceeded analysts' consensus (i.e., when SURP ≥ 0) and 0 otherwise.

Table 2 Variable description

Variable names	Descriptions
CAR[0,+3]	Cumulative abnormal return over today, and the next 3 days after the MD&A report release date
PRIOR_CAR	Cumulative abnormal return over the past ten days before the MD&A report release date
ROA	Net profit in the current quarter divided by total assets.
FUTROA	The average of ROA in the following four quarters after current quarter.
SDROA	The standard deviation of ROA in the following four quarters after current quarter.
SIZE	The natural logarithm of current quarter sales
SURP	The difference between actual net profit and the most recent forecasted analyst consensus of net profit before earnings announcement, normalized by the most recent forecasted analyst consensus.
EPS_SURP	The difference between actual earning per share (EPS) and the most recent forecasted analyst consensus of EPS before earnings announcement, normalized by the most recent forecasted EPS by

	analyst consensus.
BEAT	Dummy variable equal to 1 if announced net profit in the current quarter exceeded analysts' consensus
LOSS	Dummy variable equal to 1 if earnings are negative and 0 otherwise.
NETOPT	The difference between the positive words minus negative words scaled by total words in the MD&A documents.
DET_FS	Dummy variable equal to 1 if there is explanation about balance sheet and cash flow statement in the MD&A documents and 0 otherwise.
NONREC_POS	Dummy variable equal to 1 if there is gain form extraordinary item in the current quarter and is 0 otherwise.
NONREC_NEG	Dummy variable equal to 1 if there is loss form extraordinary item in the current quarter and is 0 otherwise.
PM	Net profit in current quarter divided by sales.
AT	Sales in current quarter divided by total asset at the end of current quarter.
DA	Total liabilities divided by total assets at the end of the current quarter.
BM	The book value of equity divided by market value of equity at the end of the current quarter.

Chapter 4
Research methods and models

Textual analysis methodology

- **Dictionary-based approach**

Another application of NLP methodology is called "dictionary-based approach". Researchers can examine the tone of particular document by counting words that have positive and negative sentiment within the document. By doing so, researchers need to identify dictionary that will be used to create a word lists. There are numbers of dictionaries that are popular among pioneer in this field including Harvard IV-4 and Diction. However, two of them are general English language linguistic dictionaries rather than dictionaries that are specific to the domain of financial document. Henry (2006), and Loughran and McDonald (2011) both argue that most of the negative word counts according to the Harvard and Diction dictionaries are attributable to words that are typically not negative in a financial context. They did suggest their own list of specific words for the domain of financial disclosure (FD and LM negative word list respectively). Jegadeesh and Wu (2013) applying LM negative word list and find that tone of a policy announcement on the releasing day of the Federal Open Market Committee (FOMC) minute meeting is correlated with significantly higher stock market volatilities. Although, Loughran and McDonald negative word list becomes a predominant in textual analysis for finance, some literatures combine several dictionaries for examining tone of financial document in order to compare a predictive power of market response. For example, Davis, Matsumoto, and Zhang (2011) use three different wordlists i.e.

DICTION, Henry (2006), and Loughran and McDonald (2011) wordlist to examine the effect of managerial “Style” on the tone of earnings conference calls.

Beside the choice of dictionary, another way to improve the accuracy of dictionary based approach is a term weighting technique. Since raw count of words will vary depend on the length of the document, researcher try to analyze sentiment words as a percentage of number of negative words to the total word-count in a document. For example, Loughran and McDonald (2011) use two weighting schemes, a simple proportional weighting and one that weights each word inversely proportional to its document frequency. Since, equal weighting would not be appropriate, regarding the fact that some words have more affect to investor than others. Although, Henry and Leone (2010) investigate all 8-K filings that include an earnings release and conclude that equal weighting of word occurrences is more intuitive, easier to implement, and more amenable to replication, Jegadeesh and Wu (2013) suggest the new term weight based approach provides a more reliable measure of document tone. They assign weights for each word based on how the market has reacted to them in the past.

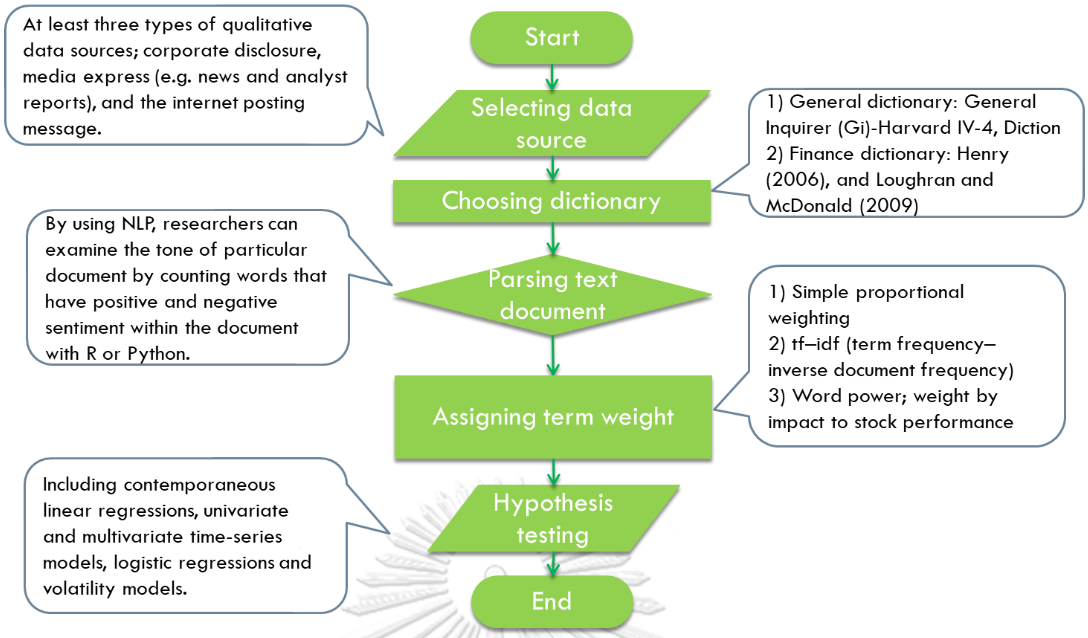

Figure 5 Flow diagram of process in extracting textual sentiment

In order to compute textual sentiment, one important issue is the choice of dictionary to match the tone of each word. I chose Loughran and McDonald (LM) Master Dictionary as my primary word list. This dictionary is made especially for financial document analysis, and it has recently become the most popular dictionary for research in finance. Their strategy to create the matching would be to let the data empirically determine the most impactful words from all 10-k during 1994-2008 and updating up to the current year. All tokens with a frequency count of 100 or more and with identifiable words are added to the dictionary. Most importantly, they could find significant relations between their word lists and file date returns, trading volume, and subsequent return volatility. Apart from positive (Fin-Pos) and negative (Fin-Neg) word lists, they propose other word lists such as litigation, uncertainty, and constraint word lists, etc.

- **Readability score**

Readability in financial disclosure, in 1998, U.S. Security Execute Commission (SEC) issued new plain English disclosure guidelines that encouraged the use of plain English in the drafting and formatting of all financial disclosure by domestic and foreign issuers. Many researchers are interested in the effect of this policy and try to measure level of readability by using Fog index. The Fog index gives the number of years of education that your reader hypothetically needs to understand the paragraph or text. A Fog index's score criteria come from the length of the sentences and number of complex words (word that have more than 2 syllables). The most prominent among these papers are Li and economics (2008) who finds that firms with lower earnings tend to file annual reports that are more difficult to read; an increase in earnings from the previous year also results in annual reports that are easier to read compared with previous year's reports. But Loughran, McDonald and Yun (2013) argue that fog index is questionable for its ability to capture a level of readability in financial market, since many complex words determined by fog index are easily understand by participants in financial market, and some larger companies that operate in diversified sector across countries tend to have more content to disclose than smaller domestic company. They suggest using natural log of gross 10-K file size available on the SEC's website due to easy to obtain, does not require problematic parsing of 10-K, less prone to measurement error and allow straightforward replication. Moreover, they also find that larger 10-Ks are significantly associated with high return volatility, earnings forecast errors, and earnings forecast dispersion, after controlling for other variables such as firm size, book-to-market, past volatility, industry effects, and prior stock performance.

I begin my readability analysis by examining the "textstat" package in Python, it calculates the statistic about readability and complexity in documents. Starting from basic statistic such as word count, syllable count and sentence count to the readability score such as The Flesch Reading Ease, Gunning FOG, SMOG Index and Coleman-Liau Index. I chose Flesch-Kincaid (FK) Grade Level to indicate readability level of Thai firms' financial disclosures because it is widely used and easy to interpret. Where the formula is as follow,

FK Grade Level = 0.39*(words/sentence) +11.8*(syllables/word) –15.59

The result of the Flesch-Kincaid grade level equal to 6 could be interpreted such that the person who has education for 6 years (a sixth-grade level) could be able to understand the context in this document.

- **Topic modeling**

Apart from the textual sentiment, thematic is the technique that can be used to classify common themes in documents or simply identify themes within a corpus of documents. While Latent Semantic Analysis (LSA) uses singular value decomposition to identify an orthogonal basis within the dimensionality constraint, Latent Dirichlet Allocation (LDA) uses a Bayesian model that views the documents as a mixture of latent topics. Still within the "bag of words" realm are techniques that can be used to classify common themes in documents or simply identify themes within a corpus of documents. Broadly, these techniques, like most, are attempting to reduce the dimensionality of the term-document matrix, in this case based on each word's relation to latent variables. After turning texts to term-document matrix, LDA assumed that selected documents share the same set of topics, but each document contains these topics in different fraction. LDA would estimate the distribution of topics then it will randomly choose a word from the matching distribution over the dictionary. Consequently, All text in the collection of documents are inserted to the model, while the hidden topic structure such as topic distributions per document, and word proportion in each topic are estimated by LDA. Only requirement that researchers need to specify is the number of topics they wish the model to discover. However, by choosing insufficient topics can produce results that are too general, while selecting numerous topics can lead to too detailed topics. Once number of topic is indicated, the LDA gives the probabilities of words being used in a topic and provides the distribution of those topics across the documents.

To examine uses of topic model in finance Hansen, McMahon and Prat (2015) make a methodological contribution by introducing Latent Dirichlet Allocation (LDA), invented by Blei, Ng, and Jordan (2003), to the economics literature. They use the set of transcripts from the tenure of Alan Greenspan - August 1987 through January 2006, inclusive, a total of 149 meetings. The most striking results are that meetings become more formal and scripted; more quantitatively-oriented; and that the amount of interjections in the debate in FOMC2 declines remarkably. Huang et al. (2018) provide one of the first applications of this method in accounting and finance, using the technique to examine the topical differences between conference call content and subsequent analyst reports. Whereas the traditional use of announcement returns made it difficult to separate out the amount of incremental information actually provided by the analysts, by comparing topical differences the authors are able to isolate the value added of analyst reports. They document that analysts provide significant and differentiated information beyond that contained in the conference call. Wu (2016) used machine learning methods on firm-level textual disclosures. He fits the LDA algorithm with 20 topics on the collection of 19,771 disclosures to identify shocks into a unique, hand-built network of firm-level supply chain connections to empirically quantify how these localized shocks affect remote firms along the chains. Surprisingly, contrary to prediction by typical network theories, these firm-specific shocks impact the revenue of firms even up to 4 connections away from the origins.

In this paper, I applied Latent Dirichlet Allocation (LDA). I used the Python tutorial code on topic modeling provided by Hansen, McMahon and Prat (2014). As for parameter settings, I follow similar specification used by Hansen and McMahon (2016). Specifically, the model is estimated at the sentence level. I have to break each document into line of sentence, using full stop as a cutoff point. After that I remove all the tables, web links, addresses, notes, special characters, letter footnotes and signatures. I also need to reduce large dimensionality of the matrix. Therefore, the important step involves stemming by removing the end of words and counting only stems, for example, the term like development/ developer/ developed becomes ‘develop’. So this stemming process attempts to group words that are grammatically different but thematically identical. The popular stemmer is the Porter algorithm. Then I plot the term frequency–inverse document frequency (tf-idf) ranking, one of the most popular term-weighting schemes, which is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus. I indicate a reasonable cutoff to be 15,000 words. The first step in estimation is to initialize a model via LDA in choosing number of topics. I choose 20 topics, since 30 topics will be too detailed and 10 topics too broad. By using a collapsed Gibb sampling of Griffiths and Steyvers (2004), I get topic allocation for every iteration of the chain and I draw 10 samples from points in the chain that are thinned by setting an interval of 50. The final topic allocation is given by taking the average of the best performing 10 samples.

Hypothesis development and economic methods

- **Textual sentiment and future firm performance**

To reduce information asymmetry, managements try to disclose their firms' current and trend of future performance to investors through various kinds of channel. MD&A documents are among the most prominent and informative used by Thai managements to normally release information about earning announcement, particularly the provision of detailed income statements. However, MD&A documents are not limited to quantitative information. Prior research also demonstrated that there is incremental information content in managers' qualitative disclosures. Therefore, my first hypothesis is as follows:

Hypothesis 1: Ceteris paribus, Net optimistic language in MD&A documents of Thai listed firms is positively associated with future firm performance.

My first hypothesis is predicated on the assumption that managers use textual sentiment in MD&A documents to communicate truthful, value-relevant information to investors. However, there are a number of options in measuring manager's earning press release language. From previous studies, Loughran and McDonald (2011) used percentage of negative words to total words. Davis et al. (2011) used the different between percentage of optimistic and pessimistic words to total words, and Huang et al. (2011) used abnormal positive tone, measured as the residuals from the annual cross-sectional regression model. The dependent variable in the future firm performances – finance literature is typically some type of firm-level or market-level performance measure such as future earnings e.g. Li (2010), Demers and Vega (2011), Huang et al. (2011), future earnings changes e.g. Li (2006), Li (2010), future returns on assets e.g. Davis et al. (2011) and future cash flows e.g. Huang et al. (2011).

I have panel data with 24 earning quarters and 49 firms sample (24 rows x 49 columns). The most common approach has been to employ the panel regression model. The most common approach has been to employ the panel regression model. I chose firm fixed-effect (FE), because I believe that each firm has specific characteristic (industry, length of document, readability score) which is not time invariant. When using FE we assume that something within the individual may impact or bias the predictor or outcome variables and we need to control for this. Furthermore, I took the Hausman test. The result shows that error term and the constant is not correlated with the regressor and I should use fixed effect.

In this paper, I will use return on assets (ROA) for the four quarters subsequent to the earnings press release date as my dependent variable.

My independent variable (x) is NETOPT, calculated by subtract the percentage of “LM Fin-Neg words” to total words in the MD&A documents to the percentage of “LM Fin-Pos words” to total words in the MD&A documents. The initial controlled variables include natural logarithm of sales in current quarter (LOGREV), ROA in current quarter (ROA) and standard deviation of ROA over the four quarters subsequent to the current quarter (SDROA) and other control variables as already mentioned in the previous section. Beside tone in overall document, the topic modeling allows me to identify the proportion of key themes and tone within each document. It is interesting to investigate which topics and tones that managements mention in order to communicate information about future firms’ performance to investors through firms’ financial disclosure. The function of signal about future firms’ performance made by managements could be express as following equation

$$\textbf{FUTUROA} = \boldsymbol{f\ \{(topic_1, topic_2, \ldots, topic_n), (tone_1, tone_2, \ldots, tone_n)\}}$$

- **Textual sentiment and stock market return**

Form hypothesis 1, if managements truthfully disclose information about firms’ performance in textual data in MD&A report, investors might react, or ignore to this information. However, I assume that everyone is rationale investor, and they should response to this information consequently while the speed of access and incorporate this information to stock price valuation is inconclusive. I further assume that investors react to information between number and text separately, only net optimistic tone or change in net optimistic tone from previous quarter are my explanatory variables in this study.

Researchers using the event study (MacKinlay 1997) methodology have also defined the dependent variable as the cumulative abnormal return (CAR) (or event period excess return) over some event window. Engelberg (2008), Feldman et al. (2008), Henry (2008), Henry and Leone (2009), Doran et al. (2010), Davis et al. (2011), Davis and Tama-Sweet (2011), Demers and Vega (2011), Huang et al. (2011), Loughran and McDonald (2011a, 2011b), Davis et al. (2012), Engelberg et al. (2012), Jegadeesh and Wu (2012), and Price et al. (2012) all employ the standard event study methodology to examine the extent to which sentiment in corporate disclosures (or news articles about disclosures) impacts on firms' cumulative abnormal returns around the 'event' or during a post-event period. Appraisal of the event's impact requires a measure of the abnormal return. There are two common choices for modeling the normal return, the constant mean return model and market model. The market model is a statistical model which relates the return of any given security to the return of market portfolio. By removing the portion of the return that is related to variation in the market's return, the variance of the abnormal return is reduced. This in turn can lead to increased ability to detect event effects. Therefore, I decided to apply market model to calculate the normal return, which is the product of daily SET index return and 360 day beta (prior the earning announcements were made) of each individual stock in my sample. Therefore, my second hypothesis follows:

Hypothesis 2: Ceteris paribus, the unexpected level of net optimistic language in MD&A disclosures is positively associated with market returns around the announcement date.

My second hypothesis is based on the assumption that investors will react to the tone of managers in MD&A documents. As a result, investors should buy the stocks that showed higher level of the different between percentage of optimistic and pessimistic words to total words, and should sell the stocks that showed lower level of the different between percentage of optimistic and pessimistic words to total words. However, the timing of impact from managers' tone in MD&A documents is still unknown. The qualitative information has been proved to have longer time to digest than quantitative data. Moreover, it is possible that the news or analyst reports could travel slowly in emerging market like Thai stock market.

In this paper, I will use cumulative abnormal return (CAR) as my dependent variable. My independent variable (x) is percentage change in NETOPT from previous quarter as a proxy of surprise in management tones in the MD&A documents. The initial controlled variables include surprise in earning per share (EPS_SURP), whether the actual earning beats consensus earning (BEAT), size of firm represent by log of revenue (SIZE) and other control variables as already mentioned in the previous section. Beside tone in overall document, the topic modeling allows me to identify the proportion of key themes and tone within each document. It is interesting to investigate which topics and tones that investor response to information in firms' financial disclosure. The function of market response to the signal form managements in MD&A reports could be express as following equation

$$\mathbf{CAR} = \boldsymbol{f\ \{(topic_1,\ topic_2,\ \ldots,\ topic_n),\ (tone_1\ ,tone_2\ ,\ \ldots,\ tone_n)\}}$$

- **Stock market volatility**

Researchers pay more attention to stock return volatility recently, because it indicates the level of dispersion between actual price and theoretically stock price, which is the proxy of risk in investment in stock. Moreover, it is one of major components in predicting future stock price, and it is used in pricing in many financial derivative products. Volatility could be measured by calculating the standard deviation or sum of square different between actual price and its mean divided by number of observation, volatility shows two stylized fact that are volatility clustering and mean reverting, which violate random walk assumption in financial time series, stated that price movements are independently and identically distributed. Therefore, previous literatures attempted to introduce new models that overcome this limitation.

The ARCH class models are now recognize as the most famous techniques for study about volatility in financial market. Specifically, the ARCH process imposes an autoregressive structure on the conditional variance that permits volatility shocks to persist over time . It can therefore allow for volatility clustering. ARCH class models make use of sample standard deviations but formulate the conditional variance, h_t, of time series via Maximum Likelihood (ML) procedure. The first example of an ARCH model is the ARCH (q) of Engle (1982) where h_t is a function of lagged past square residuals. In GARCH (p,q) additional dependencies are permitted on p lags of past realizations of the variance. I employ EGARCH, because it ensures that the conditional variance is positive and allows for the asymmetric response of the volatility to good and bad news. However, the extreme case of highly positive or negative tone could make stock prices more fluctuate than usual. I employed EGARCH (1,1) model, one of the foremost techniques for modelling volatility in financial markets, with individual SET50 stock data during Jan 2012 – Mar 2019. Therefore, my third hypothesis follows:

Hypothesis 3: Ceteris paribus, Net optimistic language in MD&A documents of Thai listed firms is negatively associated with stock return volatilities.

The main variable of interest is the variable $tone_t$, which contains the tone of managements within MD&A reports. The variables $tone_t$ are equal to zero on days when the listed firms do not announce explanations about their financial performances, given that there is no data. I also include additional control variables word_count as a proxy for the quantity of information, which listed firm provide to investors. I set word_count to 0 on days without MD&A announcements—since there is no MD&A report on such days. My hypotheses are that more positive tone and more quantity of information lower volatility, i.e., $\delta_{tone} < 0$ and $\delta_{word} < 0$. The latter hypothesis is based on the idea that relatively investors will react to bad news more aggressively. Moreover, if managements try to provide more explanation about firm performance in financial disclosure, investors would be easier to expect future firms' profitability. In line with this reasoning, we would expect to see lower stock price volatility. The model is estimated via maximum likelihood.

In order to distinguish the effect of positive and negative sentiment on the volatility listed firms in SET50. I insert $POSITIVE_t$ and $NEGATIVE_t$ variables separately to the variance equation in GARCH model. Most importantly, the expected signs of parameters are $\delta_{positive} < 0$ and/or $\delta_{negative} > 0$. This can be interpreted that the relatively negative market sentiment will lead to the high volatility. On the contrary, more positive market sentiment will lower the market volatility. Unless separate the overall tone to positive and negative, anything else is the same as I mentioned in the previous model.

Chapter 5
Analysis and discussion of results

Result from textual analysis

I programed in Python to count words that are in LM Fin-Pos list, and found that there were 33 positive words per document on average. The top frequent positive words are such as “gained”, “improved”, “strong”, “stable” and “better” (Figure 4). For LM Fin-Neg list, I found that there were 41 negative words per document on average. The top frequent negative words are such as “loss”, “impairment”, “dropped”, “declined”, “restated” and “shutdown” (Figure 4). The average number of negative words slightly exceeds the average number of positive words because there are more negative words in LM Fin-Neg list (2356 words) than positive words in LM Fin-Pos list (536 words). More importantly, I found that most of the top negative words, used in Thai listed firm’s MD&A, are the same as those Loughran and McDonald found in their work. This implies that at least Thai listed firms’ management wrote financial disclosure documents using word lists similar to those listed firm managers in the U.S. stock market.

Negative words **Positive words**

Figure 6 Top most frequent LM Fin-Pos and Fin-Neg in Thai MD&A

Particularly, the results from LDA in Table 3 show that LDA model could properly identify topics in MD&A disclosures from Thai listed companies. For example, the most frequent related words such as “price,” “oil,” “product,” “crude,” “spread,” and “fuel” in a topic of resources industry suggests that this topic is related to “energy price.” Similarly, the most frequent words such as “hotel,” “growth,” “Thailand,” “segment,” and “retail,” appear in a topic of services industry suggests that this topic relates to “tourism”. I also find that LDA effectively uncovers general topics related to income statement; balance sheet, economy, as well as cash flow topics.

Table 3 Top 10 most frequent words in 20 topics from LDA model

Topic Label	Top 10 words									
Expense (T0)	expens	cost	oper	depreci	administr	properti	equip	amort	relat	fee
	0.176	0.154	0.053	0.038	0.026	0.023	0.022	0.022	0.022	0.022
Oil price (T1)	price	oil	product	crude	spread	market	fuel	demand	suppli	ineri
	0.093	0.085	0.056	0.048	0.034	0.029	0.026	0.025	0.024	0.024
Tourist (T2)	hotel	growth	thailand	segment	retail	core	bangkok	grew	perform	trade
	0.043	0.038	0.037	0.029	0.019	0.016	0.016	0.014	0.014	0.013
Financial statement (T3)	compani	financi	manag	limit	subsidiari	statement	public	risk	consolid	posit
	0.198	0.077	0.059	0.055	0.05	0.044	0.04	0.032	0.02	0.018
Sale volume (T4)	sale	volum	per	perform	averag	product	barrel	compar	price	ga
	0.111	0.069	0.068	0.048	0.04	0.038	0.03	0.025	0.021	0.018
Balance sheet (T5)	total	asset	ratio	equiti	current	liabil	debt	account	time	financi
	0.12	0.113	0.065	0.064	0.059	0.059	0.037	0.033	0.032	0.027
Net profit (T6)	profit	incom	net	tax	margin	interest	ebitda	gross	earn	oper
	0.17	0.16	0.124	0.067	0.063	0.054	0.042	0.031	0.028	0.028
Term rate (T7)	rate	will	billion	agreement	term	remain	ore	import	fix	may
	0.091	0.043	0.028	0.025	0.024	0.02	0.019	0.019	0.017	0.015
Economy (T8)	growth	market	continu	econom	industri	demand	economi	global	domest	china
	0.049	0.044	0.04	0.026	0.025	0.022	0.022	0.021	0.02	0.019
Compare period (T9)	period	end	last	compar	decreas	previou	month	respect	day	ship
	0.164	0.119	0.086	0.079	0.069	0.06	0.047	0.028	0.025	0.018
Extra item (T10)	loss	gain	amount	exchang	foreign	thai	total	result	impair	currenc
	0.104	0.064	0.063	0.049	0.041	0.034	0.028	0.028	0.027	0.017
Project expansion (T11)	project	addit	plan	capac	construct	rate	complet	develop	approxim	expans
	0.092	0.043	0.029	0.024	0.022	0.02	0.018	0.017	0.015	0.015
Service revenue (T12)	revenu	servic	sale	satellit	network	media	total	good	mobil	internet
	0.224	0.149	0.086	0.021	0.02	0.017	0.015	0.015	0.013	0.011
Investment (T13)	share	group	oper	corpor	per	result	valu	line	total	base
	0.097	0.077	0.072	0.043	0.037	0.033	0.031	0.03	0.025	0.023
Bank customer (T14)	custom	capit	bank	develop	offer	enhanc	channel	strategi	promot	program
	0.054	0.027	0.025	0.016	0.014	0.013	0.013	0.012	0.012	0.012
Cash flow (T15)	invest	cash	net	activ	use	oper	payment	financ	dividend	receiv
	0.138	0.108	0.088	0.052	0.045	0.037	0.033	0.03	0.03	0.029
Bank loan (T16)	loan	bank	deposit	interest	fund	account	allow	npl	market	secur
	0.132	0.052	0.037	0.026	0.023	0.021	0.019	0.019	0.018	0.017
Power plant (T17)	busi	power	plant	result	gener	unit	oper	coal	electr	energi
	0.207	0.04	0.04	0.039	0.038	0.024	0.023	0.023	0.02	0.019
Improvement (T18)	improv	product	materi	yoy	futur	impact	chain	can	control	better
	0.038	0.027	0.017	0.016	0.015	0.014	0.013	0.012	0.012	0.011
Explanation (T19)	due	decreas	mainli	higher	lower	asresult	declin	drop	rose	follow
	0.163	0.145	0.102	0.087	0.067	0.04	0.037	0.028	0.024	0.023

In order to explore about topic proportion, the 20 topics classified by LDA model are still too many to analyze. I need new tool to segment these 20 topics to reduce topic dimension. I then use tree-based hierarchical diagram or 'dendrogram' in the figure below to condense the topics, used by management in financial disclosure documents. The main purpose is to see how these 20 topics are related to each other. The closer at which any of the two topics is linked, the more similar their words using patterns are. Reading the figure from the bottom-up shows information about which clusters are merged first at the lowest height. It can be seen that topic 6 and 0 are basically in the same category about firm performance that are merged first due to most similarity in the word usage. Then topic 15, 19 belongs to tonal words or sentiment change in this cluster. Topics 1, 9, and 4 are considered another clustering set about energy prices and sales volume in the energy market. Then, topics 7, 8 and 18 are essentially about economic growth condition. This figure illustrates graphically an important feature of the MD&A documents. That is, firm financial performance is a complicated explanation and the management needs to mention number of factors when communicating appropriate information to investors. When viewed this dendrogram from the top down, I can see that topics 2 and 12 are clearly about service business, and topics 14 and 16 are about banking business. Moreover topic 1, 4 and 9 are about energy price, while topics 13 and 17 are about power generation. The management then informs about financial status (T3, T5 T10 and T11) and assessing its financial performance relative to four main clusters on expense, net profit, cash flow and explanation (T0, T6 T15 and T19). Last but not least, managements describe about the real economic growth condition as the external factor (T7, T8, and T18).

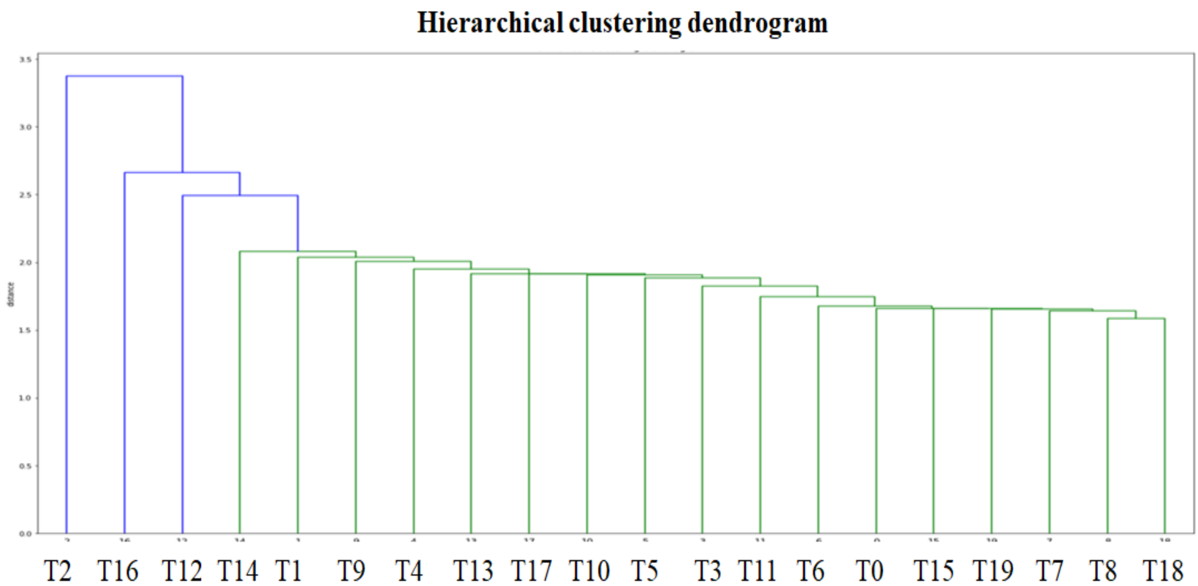

Figure 7 Tree-based hierarchical diagram

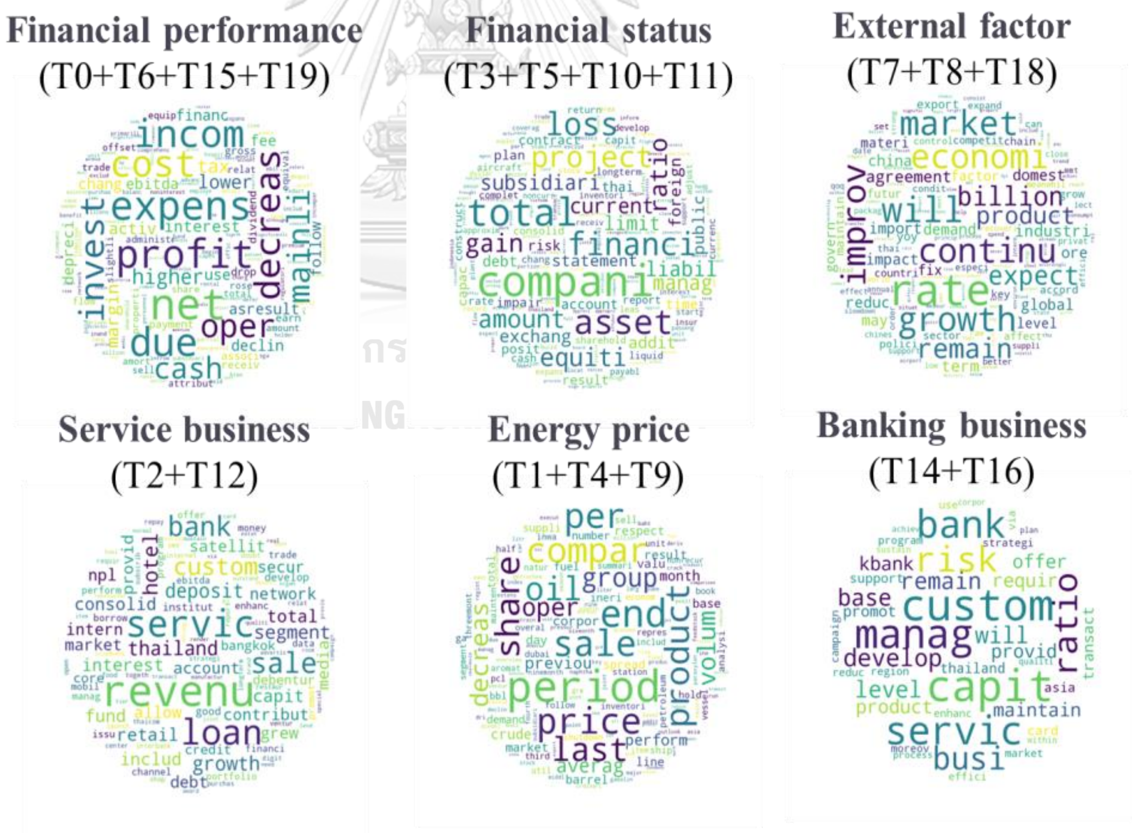

Figure 8 top most frequent words in reduced form topics

Figure 9 illustrates how the topic discussions in MD&A document evolve between quarter 1/2012 and 4/2018. I find that each topic is quite persistent during the time of study. The proportion of discussion on financial performance (FP) and financial status (FS) are averaging at 23% and 21% respectively, while role of the external factor (EF) is around 13% per document. After all, I combine other topics: banking business, service business, energy price and power generation together and label as "industry specific" (IS) topic, which is the largest topic proportion in MD&A document. The proportion of discussion is around 42%, with the highest standard deviation of 11%. However, the overall proportion of discussion topic shows low variation across time compare to SET Index. Therefore, this can imply that using the proportion of discussion topic captured by LDA model alone may not reflect the variation of stock market response.

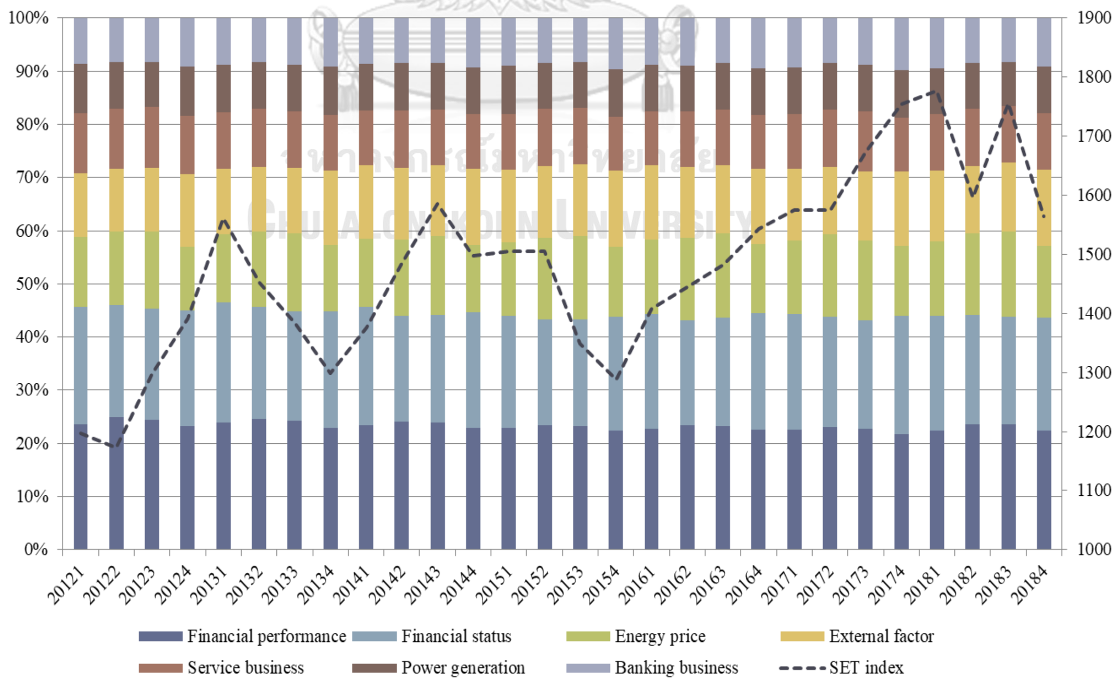

Figure 9 Average topic proportion in MD&A documents across time

Figure 10 presents visual evidence of a reliable capability of LDA in order to capture the key topics from MD&A documents in various industries. In the banking industry, for example, management and analyst discussions are devoted primarily to the topics of “banking business” mostly in “loans”, “deposits” “NPL” and “capital ratio.” The discussion of financial performance topics increased substantially in Q1/2016 around 4.8%, after the net profit as a whole industry went up from both interest and non-interest income. In the resource industry, for example, management and analyst discussions are devoted primarily to the topics of “energy price” mostly in “crude”, “oil price” “spread” and “production.” Not surprisingly, after the industry has been hit hard by the current downtrend of low energy prices in early of 2015, there is a decrease in discussions of the topic financial status about 3.6% in Q3/2015. Moreover in Technology industry, the management discussions are devoted primarily to the topics of “service business” mostly in “revenue”, “satellite”, and “network”, “mobile” and “internet”. It shows that external factor proportion of topic increased dramatically during Q4/2015 after 4G service license auction held in November 2015.

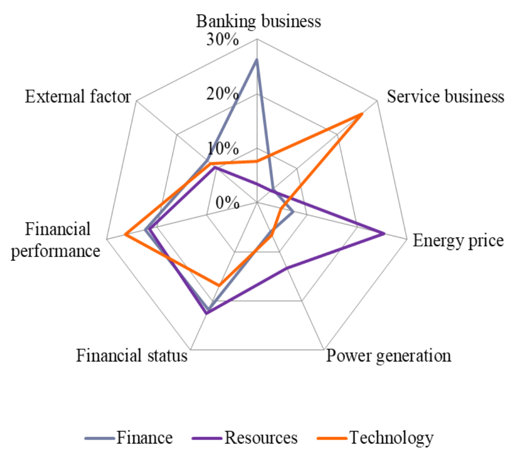

Figure 10 Average Topic proportion in MD&A documents across industries

In this section, I investigate the management's tones of each topic whether they have information about future firm performance and how investors react to them. Beside the proportion of topic from LDA, I will focus on the net tone specifically at four major topics, i.e. financial performance (FP), financial status (FS), external factor (EF) and industry specific (IS), respectively. By using the word list developed by Loughran and McDonald (2011), I then use automated dictionary method by counting the number of the positive and negative words at the sentence level of every MD&A documents, and normalizing by total number of stems in each topic. In addition, I also adopt uncertainty word list from Loughran and McDonald in order to monitor the level of ambiguity in financial disclosure. The total sentence of the MD&A of 47 firms over the past 7 year is 163,352. I find that managements write MD&A documents with negative tone in every topic, except industry specific topic in some industries, and the financial performance topic shows the lowest net tone (as shown in figure 11). Moreover, I find that the topics that contain relatively more uncertainty words than others are external factor and financial status topic (as shown in figure 12). Therefore, it can imply that managements normally use tonal words in FP topic to express about sentiment, while use uncertainty words in EF and FS topic when the condition is unclear.

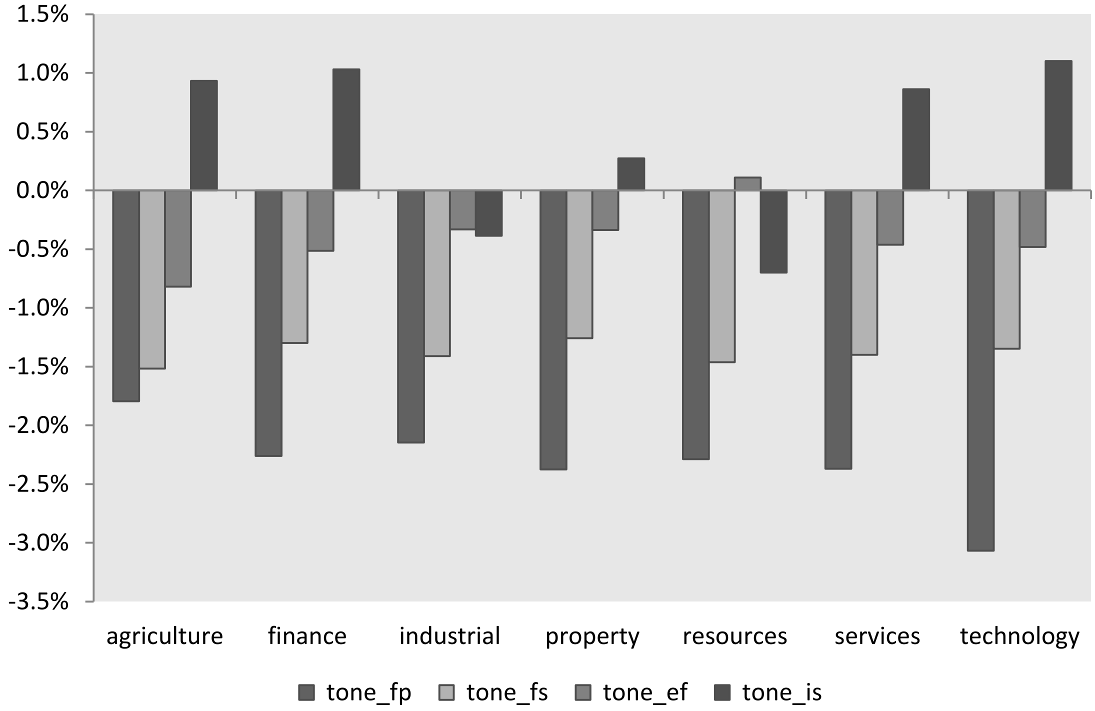

Figure 11 Average tones of each topic across industries

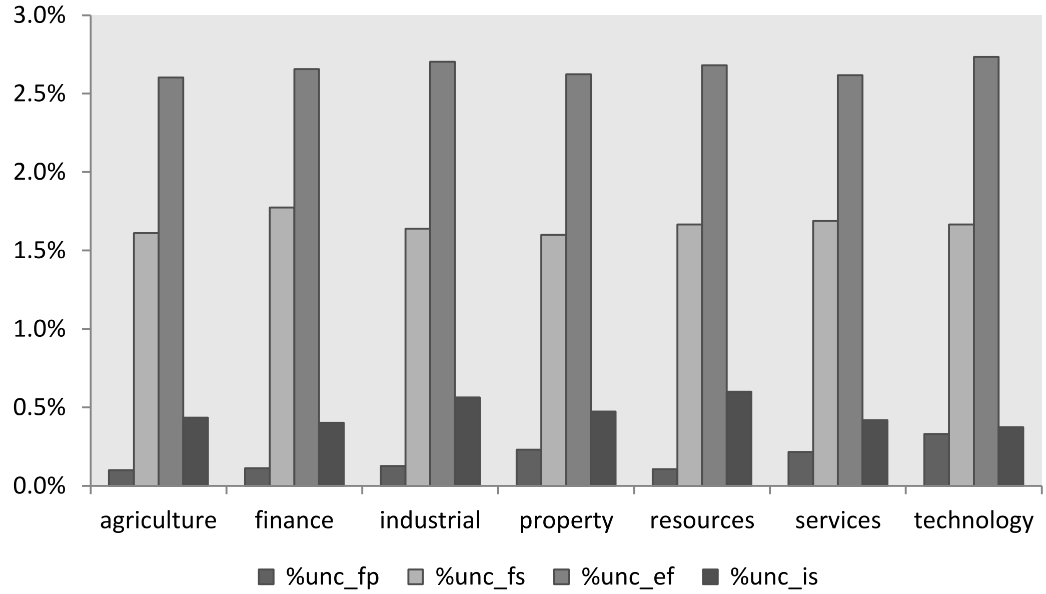

Figure 12 Uncertainty words of each topic across industries

To evaluate document complexity, I calculated readability measures i.e. Flesch Reading Ease Score and Flesch-Kincaid Grade Level by applying "textstat" package in Python program. I found that the overall Flesch Reading Ease Score and Flesch-Kincaid Grade Level in my sample are 12.6 and 21.5 respectively, which identified that Thai listed firms' MD&A reports are very difficult to read. A number of reasons could explain, first there are many listed firms in my sample that segment in healthcare and resources sectors, which their businesses are complicate and normally use technical terms i.e. "Olefins", "Propylene", "Oxide", "Polyols" and "Methy". Second, managements write MD&A in Thai-English style, which might include Thai words, for example, name of province ("Nakorn Ratchasima", "Ratchaburi", "Rayong"), and name of subsidiaries ("Dusit", "Bumrungrad", "Ramkhamhaeng"). Many of these words are not in English dictionary; as a result computer algorithm will see these as the complex words, which drive the readability index to be high. Therefore, in order to avoid measurement errors, I propose to use length of MD&A report (e.g. number of words, number of sentence) as the readability level in this study.

Result from economics model

Quantifying and plotting the information about financial data and MD&A reports allows us to visualize its content and how this has evolved over time.

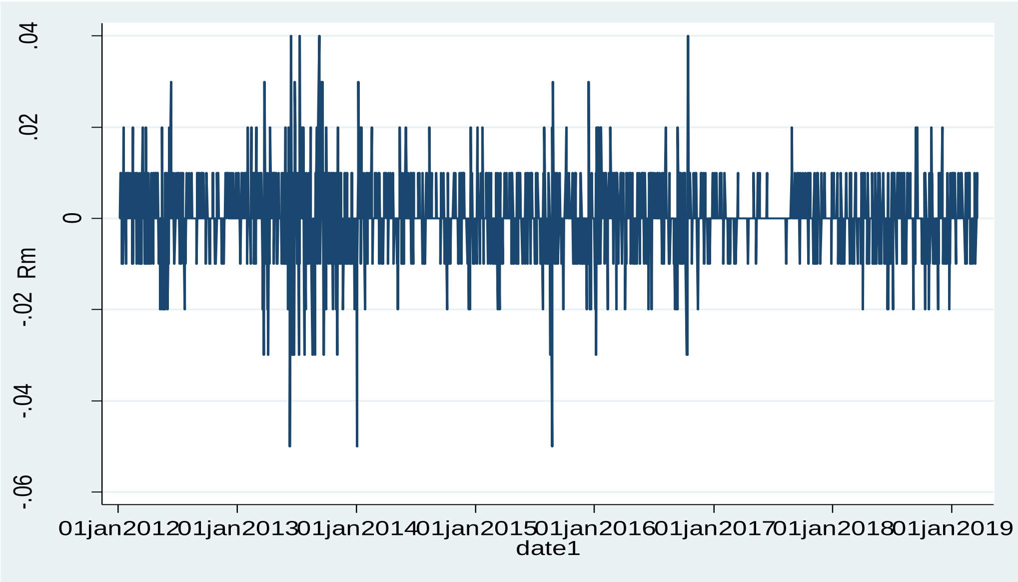

Figure 13 Daily Stock Market Returns

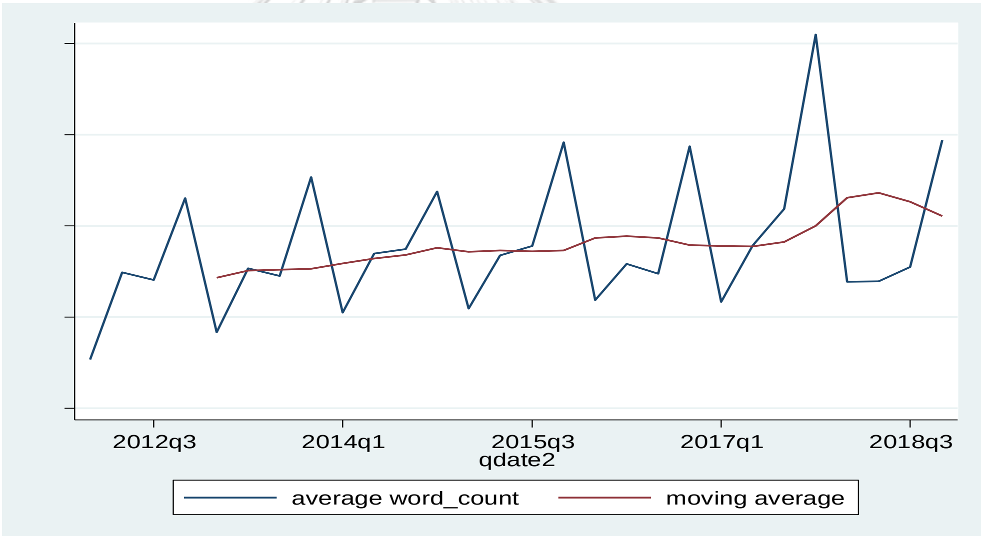

Figure 14 Length of MD&A reports

Figure 13 plots stock market return in percentage change from Jan 2012 to Mar 2019. The graph shows the evidence that volatility clustering persists in the Stock Exchange of Thailand during the time of study. The daily stock return is higher (lower) volatile follow by highly (lowly) volatile period, where the daily maximum drawdown is -5.37 percent while daily maximum gain is 4.48 percent. However, the average daily price change is around 0.00 percent. Figure 14 plots the average number of words per MD&A report of 25 listed firms in SET50, along with a moving average that facilitates the visualization of medium-term trends. Overall, MD&A reports have become considerably longer—from around 2,700 words at the beginning of our sample to around 3,100 words at the end of the sample, with a peak of around 4,000 words in q4 2017.

Descriptive statistic

Table 4 presents descriptive statistics for all accounting, financial market, and textual-analysis variables. Since these 47 firms' distribution of revenue is highly skewed. I replace natural logarithm of revenue in the model instead as proxy of firms' size. FUTROA and ROA are very close, which implied that future firm performance is quite persistent. Overall, sample firms financial position look strong, only a few quarter showing negative sign for earning and more than one half of their earning each quarter and beat analyst consensus. What should be concerned is SURP, since it has high standard deviation and the gap between maximum and minimum value is extremely high. However, EPS_SURP is more stable with lower standard deviation. Managements write MD&A report quite neutral, the average of NETOPT is quite neutral at -0.003. Managements write MD&A report quite neutral in financial performance topic, the average of TONE_FP is averaging at -0.024. Moreover, they use uncertainty words in financial status topic around 1.7% of total words per sentence. Cumulative abnormal return for the next three days, CAR [0,+3], is average at -0.001, which implies that information in MD&A disclosures have small impact to stock returns in the next three days after documents were released.

Table 4 Descriptive statistics panel data for hypothesis 1&2

	Obs.	Mean	Maximum	Minimum	Std. Dev
A: Textual data					
NETOPT	1,316	-0.003	0.043	-0.041	0.012
TOPIC_FP	1,316	0.233	0.555	0.045	0.081
TOPIC_FS	1,316	0.211	0.507	0.054	0.078
TONE_FP	1,316	-0.024	0.018	-0.038	0.009
UNC_FS	1,316	0.017	0.024	0.011	0.002
B: Financial data					
CAR[0,+3]	1,316	-0.001	0.349	-0.400	0.041
PRIOR_CAR	1,316	0.000	0.364	-0.367	0.054
C: Accounting data					
FUTROA	1,316	0.055	0.361	-0.306	0.078
ROA	1,316	0.056	0.412	-0.318	0.080
SIZE	1,316	3.864	5.875	1.827	0.826
SURP	1,316	-1.124	24.700	-577.672	21.488
BEAT	1,316	0.534	1.000	0.000	0.500
LOSS	1,316	0.854	1.000	0.000	0.354
PM	1,316	0.145	6.463	-3.657	0.369
AT	1,316	0.181	0.874	0.013	0.164
DA	1,316	0.275	0.789	0.000	0.151
BM	1,316	0.642	3.012	0.037	0.406
EPS_SURP	1,316	-0.059	76.500	-32.90	3.414

The table shows summary statistics for the textual data in listed firms' financial disclosure in Panel A, the financial data in Panel B, and the accounting data in Panel C.

Correlation statistic

Table 5 presented the correlation matrix among textual sentiment, financial data and accounting data. Some of the variables are correlated with each other. Textual sentiment is correlates to return on asset at 0.21, which indicates that managements use optimistic language to communicate current firms' performance to investors. However, the only variable that highly associated to cumulative abnormal return is BEAT, which implied that whether the actual earning per share beat analyst consensus impacts stock abnormal return.

Table 5 Correlation matrix

	NETOPT	ROA	CAR[+1,+3]	SURP	EPS_SURP	BEAT	PM	AT	DA	BM
NETOPT	1.00									
ROA	0.21	1.00								
CAR[+1,+3]	0.02	0.03	1.00							
SURP	0.06	0.09	0.03	1.00						
EPS_SURP	0.02	0.02	0.07	0.16	1.00					
BEAT	0.10	0.06	0.19	0.08	0.10	1.00				
PM	0.07	0.49	0.06	0.08	0.11	0.14	1.00			
AT	-0.02	0.08	0.02	0.01	0.08	-0.02	-0.22	1.00		
DA	-0.07	-0.19	-0.06	0.02	-0.03	-0.02	-0.16	0.16	1.00	
BM	-0.27	-0.48	0.00	-0.08	-0.04	-0.04	-0.24	0.06	0.04	1.00

Test of hypothesis 1 textual sentiment and future firms' performance

The first hypothesis (H1) predicts that net tone in MD&A reports is positively associated with future firm performance. To test H1, I used a panel regression model with firm fixed effect to explain future firm performance, based on that used in Davis, Piger, and Sedor (2011). Future firm performance (FUTROA) is measured as the average of ROA in the four quarters subsequent to the current quarter. The following model is then used to explain FUTROA:

$$FUTUROA = \beta_0 + \beta_1 ROA_{it} + \beta_2 SDROA_{it} + \beta_3 SIZE_{it} + \beta_4 SURP_{it} + \beta_5 BEAT_{it} + \beta_6 LOSS_{it} + \beta_7 DET\_FS_{it} + \beta_8 DIV\_INC_{it} + \beta_9 NONREC\_POS_{it} + \beta_{10} NONREC\_NEG_{it} + \beta_{11} PM_{it} + \beta_{12} AT_{it} + \beta_{13} DA_{it} + \beta_{14} BM_{it} + \beta_{15} NETOPT_{it} + \varepsilon_{it} \tag{1}$$

From the result of regression model for equation (1) in table 6, I used panel regression with firm fixed effect. The Wooldridge test was used to justify autocorrelation in panel data (Szarowska and Finance 2018)[1]. The overall model seems to explain the firm's future performance quite well, since the overall R-square is at 0.7613. More importantly, the coefficient on NETOPT is positive and significant, suggesting that higher values of NETOPT predict higher future firm's performance. Generally, managers use language in earnings MD&A report to communicate incremental, value-relevant information to investors and other stakeholders. For other explainable variables, the coefficient on ROA is estimated to be positive and strongly significant, consistent with prior research. The coefficient on SURP is positive and statistically significant, suggesting that earning surprise positively correlated with future firm's performance. The coefficients on PM and BM are negative and statistically significant.

---

[1] By using Wooldridge test for autocorrelation in panel data. The test result shows that I have to reject the null hypothesis and conclude that the data has serial correlation. However, this issue is generally accommodated by using robust standard error estimates for linear panel model.

Table 6 Tests of the association between NETPOT and FUTROA

Variable	Regressand: FUTROA Panel with Fixed Effect Coefficient	Robust SE	t-stat
INTERCEPT	0.1153	0.0909	1.27
ROA	0.5825***	0.1264	4.61
SDROA	-0.1730	0.2362	-0.73
SIZE	-0.0163	0.0216	-0.76
SURP	0.0001**	0.0000	2.17
BEAT	0.0019	0.0030	0.62
LOSS	0.0022	0.0044	0.50
DET_FS	-0.0224***	0.0067	-3.32
NONREC_POS	-0.0035	0.0024	-1.45
NONREC_NEG	-0.0021	0.0041	-0.51
PM	-0.0063*	0.0078	-0.81
AT	0.0543	0.0581	0.93
DA	-0.0016	0.0298	0.05
BM	-0.0217**	0.0088	-2.47
NETOPT	0.5296*	0.2654	2.00
Overall R-square	0.7613		
Sample Size	1,176		

*/**/*** denote statistical significance at 10%, 5%, and 1% levels, respectively, based on a two-tailed t-test.

- **Topic proportion and tone within each topic**

Although the first hypothesis predicts that overall tone in MD&A document has correlation with future firm performance. In order to investigate how managements mention about topic and tone could give more understanding about what information inside a financial disclosure. Since I expected that **FUTUROA = *f{(topic_is, topic_fp, topic_fs,***

topic_ef), (tone_is, tone_fp , tone_fs, tone_ef)}, I created 8 more variables according to what I found from LDA model. Moreover, I proposed to use LM uncertainty word list as alternative information beside net optimistic language. To test this hypothesis, I assess information extracted from MD&A to future firm's performance by estimating the following regression:

$$
\begin{aligned}
FUTUROA = {} & \beta_0 + \beta_1 TOPIC\_IS_{it} + \dots + \beta_4\, TOPIC\_EF_{it} + \\
& \alpha_1 TONE\_IS_{it} + \dots + \alpha_4 TONE\_EF_{it} + \\
& \gamma_1 TOPIC\_IS_{it} \times TONE\_IS_{it} + \dots + \gamma_4\, TOPIC\_EF_{it} \\
& \times TONE\_EF_{it} + Controls + \varepsilon_{it}
\end{aligned}
\qquad (2)
$$

Table 7 Tests of the association between tone and FUTROA

Dependent variable	**FUTUROA**						
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ROA	0.5824***	0.6122***	0.6038***	0.6038***	0.6119***	0.6085***	0.6038***
SDROA	-0.1694*	-0.1605	-0.1670*	-0.1670*	-0.1650*	-0.1602	-0.1429
SIZE	-0.0246*	-0.0297**	-0.0277*	-0.0277*	-0.0347*	-0.0308**	-0.0341**
SURP	0.0000*	0.0001*	0.0001**	0.0001**	0.0001*	0.0001**	0.0001**
LOSS	0.0040	0.0070	0.0065	0.0065	0.0069	0.0071	0.0074
NONREC_POS	-0.0045	-0.0045	-0.0042	-0.0042	-0.0047	-0.0036	-0.0034
NONREC_NEG	-0.0035	-0.0054	-0.0050	-0.0050	-0.0054	-0.0048	-0.0042
PM	-0.0040	-0.0048	-0.0045	-0.0045	-0.0050	-0.0054	-0.0045
AT	0.1209***	0.1189***	0.1202***	0.1202***	0.1187***	0.1207***	0.1229***
DA	0.0381**	0.0419**	0.0401**	0.0401**	0.0412*	0.0403**	0.0389**
BM	-0.0009	0.0005	0.0028	0.0028	-0.0013	0.0000	-0.0031
NETOPT	0.6089***						
TOPIC_IS		0.0080		0.0044	0.0044	0.0023	-0.0636
TOPIC_FP		-0.0792		-0.0760	0.2067	-0.0829	0.0557
TOPIC_FS		0.0000		0.0000	0.0000	0.0000	0.0000
TOPIC_EF		0.0649		0.0642	0.0030	0.0716	-0.0171
TONE_IS			-0.0861	-0.0460	-0.5958		
TONE_FP			0.4001	0.1660	-1.694**		
TONE_FS			0.1685	0.1252	0.0301		
TONE_EF			0.0120	0.0536	0.9556**		
UNC_IS						0.2403	-6.2333
UNC_FP						-1.6049	15.513***
UNC_FS						1.2878*	1.3634
UNC_EF						0.2773	0.1350
TOPIC_IS x TONE_IS					1.2035		
TOPIC_FP x TONE_FP					9.677***		
TOPIC_FS x TONE_FS					0.5360		
TOPIC_EF x TONE_EF					-14.9445***		
TOPIC_IS x UNC_IS							14.4002
TOPIC_FP x UNC_FP							-70.03***
TOPIC_FS x UNC_FS							-1.2657
TOPIC_EF x UNC_EF							3.3705
R-sq	0.7003	0.7288	0.7339	0.7236	0.7211	0.7143	0.7007

Notes. This table 7 presents coefficient estimates and t-statistic for model (1-7), */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively.

From the result of regression model for equation (2), I used panel regression with firm fixed effect, since there may have individual firm's characteristic that might be time invariant. The overall model seems to explain the future firm's performance quite well, since the overall R-square is improving after including topic proportions, tone of each topic and interaction terms between these two variables in model 2-5. More importantly, the coefficient on TONE_FP and TONE_EF are significant suggesting that the level of tone in specific topic could predict future firm's performance. The coefficient of TOPIC_FP x TONE_FP is positive and highly significant implied that generally, managers explain more about firm profitability in earnings MD&A with more positive tone, when future firms' performance shows the sign of expansion. From model 6-7, by adding uncertainty word count, I found that UNC_FS is positive and also significant suggesting that higher % of uncertainty words in FS topic predicts higher future firm's performance. For other explainable variables, the coefficient on ROA is estimated to be positive and strongly significant, consistent with prior research. The coefficient on SURP is positive and statistically significant, suggesting that earnings surprise positively correlated with future firm performance. The coefficient on SIZE is negative and statistically significant while coefficients on AT and DA are positive and statistically significant.

- **Does LDA with whole corpus outperform sectoral corpus?**

From the result of table 7, I find that there is a sign of multicollinearity among proportions of topic. I proposed to remove some proportion and tone within topics. I assume that only partial tone in FP topic and proportions of uncertainty word in FS topic have predictive power over future firm performance. To understand the interaction effect between proportion of topic and tone, I estimate the following regression:

$$FUTUROA = \beta_0 + \beta_1 TOPIC\_FP_{it} + \beta_2\, TONE\_FP_{it} + \beta_3 TOPIC\_FP_{it} \times TONE\_FP_{it} + \alpha_1 UNC\_FS_{it} + Controls + \varepsilon_{it} \tag{3}$$

Equation (3) was modified from the regression model of Equation (2) by including only interaction term between proportion of FP topic or tone within FP topic to estimates whether proportion of FP topic or tone within FP topic has information about future firm performance. I find that coefficient on the interaction terms $TOPIC\_FP_{it} \times TONE\_FP_{it}$ is positive and significant (at least at the 10% level), supporting my prediction that managements provide more proportion on FP topic with more positive net tone, when future ROA is increasing. Moreover, UNC_FS is still positive and significant, suggesting that future firm performance is positive when management disclosures are more of ambiguous language.

To further investigate, which result of LDA from whole corpus or sectoral corpus outperforms in explanation of the future performance? I divide the whole corpus into sector, then applying the entire process of LDA as same as the explanation in the previous section. Next, I put proportion of topic and measure of the tone within topics into equation (3). Table 8 reports the regression results compared between whole corpus and sectoral corpus. The overall R-square form the whole corpus LDA has a higher value at 0.7195 compare to 0.7043 from sectoral corpus. Moreover, all of the coefficients of independent variables show the right sign and statistical significance. Therefore, it can imply that LDA with the whole corpus outperforms LDA with the sectoral corpus.

Table 8 Tests result between whole corpus and sectoral corpus

	Regressand: FUTROA			
Variable	Whole corpus		Sectoral corpus	
	Coefficient	t-stat	Coefficient	t-stat
INTERCEPT	0.0428	0.80	0.0393	0.75
ROA	0.5903***	22.01	0.5950***	22.29
SIZE	-0.0246**	-1.99	-0.0192	-1.54
SURP	0.0002***	3.97	0.0002***	3.99
LOSS	0.0151***	3.81	0.0143***	3.60
PM	-0.0022	-0.61	-0.0015	-0.40
AT	0.1095***	4.44	0.1021***	4.10
DA	0.0550***	30.9	0.0507***	3.60
BM	0.0023	0.41	0.0033	0.57
TOPIC_FP	0.0432	0.48	-0.0311	-0.83
TONE_FP	-0.8824	-1.43	-0.1501	-0.65
TOPIC_FP × TONE_FP	5.6375*	1.80	0.6338	0.66
UNC_FS	2.1124***	3.16	1.7451**	2.47
Overall R-square	0.7195		0.7043	
Sample Size	1,128		1,128	

Notes. This table presents coefficient estimates and t-statistic for Equation (3), */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively, based on a two-tailed t-test.

Test of hypothesis 2 textual sentiment and stock returns

- **Event study**

According to previous studies in developed stock markets, the textual sentiment affects abnormal return only in short window period. However, the earning announcement effect in Thai stock market was still unknown. Having established that investors respond to information about topic and tone in earning disclosures, I examine the effect through stock return around the announcement date. I estimate a panel regression model which CAR (cumulative abnormal return over three day after the MD&A report release date) is dependent variable using the following model:

Table 9 Tests of market response in different event window period

Dependent variable	CAR [0,0]	CAR [-1,+1]	CAR [-2,+2]	CAR [-3,+3]	CAR [0,+1]	CAR [0,+3]	CAR [0,+5]
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SIZE	-0.0112**	0.0051	0.0094	0.0259*	0.0039	0.0278**	0.0174
EPS_SURP	-0.0004	0.0009	0.0014*	0.0018**	0.0006	0.0009	0.0018**
BEAT	0.0028**	0.0043**	0.0102***	0.0135***	0.0055***	0.0136***	0.0135***
PM	0.0029	0.0054	0.0044	0.0067	0.0049	0.0084*	0.0100
AT	0.0181	0.0311	0.0325	0.0157	0.0193	0.0152	0.0240*
DA	0.0162	0.0144	-0.0054	-0.0102	0.0112	-0.0077	0.0199
BM	0.0035	-0.0029	0.0009	0.0008	0.0018	0.009	0.0037
TOPIC_IS	0.1295**	0.1972**	0.2212*	0.2043	0.1276	0.2074*	0.1925
TOPIC_FP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
TOPIC_FS	0.1852**	0.2110*	0.3458**	0.2993	0.1327	0.2670*	0.0539
TOPIC_EF	0.2076***	0.3157***	0.3743***	0.3953**	0.2093**	0.3921***	0.2883**
TONE_IS	0.1489	-0.1717	-0.6790	-0.7013	-0.2482	-0.4845	-0.3034
TONE_FP	-0.3686	-0.1120	-0.7086	-0.2493	0.0529	0.3038	0.7842
TONE_FS	1.8521***	2.8020**	3.460**	4.0336**	1.7109*	3.9061***	3.7029**
TONE_EF	0.8157	0.2947	0.4083	0.2703	0.3391	0.4443	1.3774
TOPIC_IS x TONE_IS	-1.2910	-0.9911	-1.2831	-0.6317	-0.7471	-0.9466	-2.9982
TOPIC_FP x TONE_FP	1.5776	0.7108	3.2848	1.4564	0.0095	-0.3887	-5.1234
TOPIC_FS x TONE_FS	-9.654**	-15.909**	-18.7034**	-21.2715**	-9.4175*	-19.7957**	-18.6701**
TOPIC_EF x TONE_EF	1.0341	9.7677	16.9110**	15.5698**	6.8721*	11.8771**	7.5973
R-sq	0.0073	0.0106	0.0165	0.0153	0.0065	0.0152	0.0199

Notes. */**/*** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively

- **How investors response to tones in different topics in MD&A**

As I expected, the earning announcement effect in Thai stocks is temporary, but investors take a little more period to adjust this information to stock price valuation. Unlike developed market, the announcement effect would be existed about 2-3 days after the announcement date. I would choose the CAR[0,+3] as my dependent variable, since it shows significant and correct expected sign of coefficients. Then I ran panel regression model again and added LM positive word list and negative word list (including interaction terms) using the following model.

$$
\begin{aligned}
CAR[0,+3] = \beta_0 &+ \beta_1 TOPIC\_IS_{it} + \ldots + \beta_4 TOPIC\_EF_{it} + \\
&\alpha_1 TONE\_IS_{it} + \ldots + \alpha_4 TONE\_EF_{it} + \\
&\gamma_1 TOPIC\_IS_{it} \times TONE\_IS_{it} + \ldots + \\
&\gamma_4 TOPIC\_EF_{it} \times TONE\_EF_{it} + Controls + \varepsilon_{it}
\end{aligned} \tag{4}
$$

Table 10Tests of the association between topics and tones to CAR

Dependent variable	**CAR [0,+3]**							
Independent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SIZE	0.0311**	0.0307**	0.0294**	0.0304**	0.0292**	0.0278**	0.0298**	0.0269**
EPS_SURP	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.001
BEAT	0.0131***	.0129***	0.0135***	0.0128***	0.0133***	0.0136***	0.0136***	0.0131***
PM	0.0088*	0.0087*	0.0084*	0.0085*	0.0081*	0.0084*	0.0083*	0.0083*
AT	0.0209	0.0218	0.0176	0.0227	0.0179	0.0153	0.0239	0.0099
DA	-0.0041	-0.0037	-0.0007	-0.0012	0.0005	-0.0077	0.0041	-0.0087
BM	0.0113*	0.0114*	0.01*	0.0119*	0.0106*	0.009	0.0091	0.0083
NETOPT	-0.038							
%CH_NETOPT		0.0599						
TOPIC_IS			-0.0594		-0.0594	0.2074*	-0.1997	1.1323***
TOPIC_FP			-0.01049		-0.0032	0.2670*	-0.2146	1.0524**
TOPIC_FS			0.0000		0.0000	0.0000	0.0000	0.0000
TOPIC_EF			0.0453		0.0519	0.3921***	-0.2478	1.8187***
TONE_IS				0.0771	0.0728	0.4443		
TONE_FP				0.1437	0.1552	0.3038		
TONE_FS				0.6425	0.6322	3.906***		
TONE_EF				0.1411	0.2141	-0.4845		
POS_IS							0.3197	
POS_FP							0.5425	
POS_FS							1.2406	
POS_EF							-1.2077**	
NEG_IS								0.985
NEG_FP								-0.3679
NEG_FS								-4.8488***
NEG_EF								0.1415
TOPIC_IS x TONE_IS						-0.9466		
TOPIC_FP x TONE_FP						-0.3887		
TOPIC_FS x TONE_FS						-19.796**		
TOPIC_EF x TONE_EF						11.8771**		
TOPIC_IS x POS_IS							-1.6633	
TOPIC_IS x POS_FP							2.1707	
TOPIC_FP x POS_FS							-4.9499	
TOPIC_FS x POS_EF							1.4826	
TOPIC_IS x NEG_IS								-4.4722
TOPIC_FP x NEG_FP								1.0101
TOPIC_FS x NEG_FS								24.3067***
TOPIC_EF x NEG_EF								-11.1949**
R-sq	0.0125	0.0127	0.0107	0.0142	0.012	0.0152	0.0126	0.022

Notes. This table presents coefficient estimates and t-statistic for Equation (4), */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively, based on a two-tailed t-test.

From regression result in table 10 investors responded to information about topic and tone in earning disclosures differently. The coefficients were significant when the regression include topics, tones and their interaction terms together. The topics and tones that have impact to abnormal returns are firms' financial status and external factor, the coefficients of these interaction terms are highly significant at 99% and 95% confidence level respectively. Moreover, results in Table 10 suggest that investors place a greater weight on negative sentiment express in firms' financial status and external factor. Therefore, managements shouldn't report only firms' financial performance. My control variable's coefficient such as BEAT is positive and significantly associate with CAR in the next three days. And also firm characteristics that impact its information environment, including firm size (SIZE), profit margin (PM) and book-to-market ratio (BM).

Test of hypothesis 3 textual sentiment and stock price volatilities

Table 3 provides summary statistics for the variables that measure two types of data; a financial data and textual data (tone and readability of the MD&A reports). Looking first at financial data (Panel A), the average stock returns is approximately 0.00 percent, calculated from earning announcement 700 days, which imply that the information in MD&A reports is quite small. The evidence is also true in term of stock market return. However individual stock prices are more volatile (approximately +/- 2.8 percent) while stock market return is around +/- 6.5 percent. The percentage change of trading volume has its standard deviation at 60 percent. The textual data (Panel B), I find that, because of the long sample period under study, the tone of the press releases is close to zero on average (only slightly negative). There is a wide variation over time, with the tone ranging from -0.77 (very negative) to +1.00 (very positive). The length of MD&A report is 111 sentences or 2,898 words on average. The longest MD&A composed of 18,445 words while the shortest one composed of 74 words

Table 11 Descriptive statistics of time series data for hypothesis 3

	Obs.	Mean	Std. Dev	Maximum	Minimum
A: Financial					
Stock Return	700	-0.000	0.017	0.065	-0.066
Market Return	700	-0.000	0.007	0.028	-0.027
Ch% volume	700	0.108	0.607	2.166	-1.790
B: Textual data					
Tone	700	-0.071	0.377	1.000	-0.773
Positive	700	0.464	0.189	1.000	0.114
Negative	700	0.536	0.189	0.886	0.000
Word_count	700	2,898	2,066	18,445	74
Sentence_count	700	111	78	562	8

The table shows summary statistics for the financial data in the stock exchange of Thailand in Panel A, the tone and readability measures of MD&A reports in Panel B

In the study, I choose listed firms in SET50 as my sample, where half of them announce MD&A reports in every quarter while the other half did not. More importantly, my independent variables are classified into 2 groups, namely financial data and textual data. Financial data are return on SET index and change in trading volume in SET. The market return data enter into EGARCH (1,1) in the mean equation in the form of percentage changes while percentage changes in trading volume data is set in variance equation. The textual data in the study are the net tone, the positive sentiment and the negative sentiment. All of which is extracted from the corpus of MD&A extracted from 25 Thai listed companies. The conditional mean equation is formulated as

$$r_t = \beta_0 + \beta_m Rm_t + \mu_t \qquad (5)$$

In line with much of the related literature, I estimate the model over all business days in the sample, i.e., including days when the listed firms do not announce its quarterly financial performance via MD&A reports . Accordingly, t denotes trading days. r_t is the daily market return. My hypothesis is that individual stock prices will move the same direction with the market return, i.e., that $\beta_m > 0$. The specification of the mean equation is simple in itself, but it is also important to ensure that I identify the coefficients in the variance equation appropriately. A failure to control for all relevant factors in the mean equation will lead to larger residuals and a higher conditional variance of the disturbance, where $\mu_t \sim (0,h_t)$. I express the conditional variance h_t as;

$$\log(h_t) = \gamma_0+\gamma_1\left(\frac{\mu t-1}{\sqrt{ht-1}}\right)+\gamma_2\left(\left|\frac{\mu t-1}{\sqrt{ht-1}}\right| - \sqrt{2/\pi}\right)+\sum_{t-1}^{1} \gamma_{2+k} \log(h_{t-k}) \qquad (6)$$
$$+ \delta_{vol}vol_t + \delta_{tone}tone_t + \delta_{word}word\_count_t$$

- **Test between with and without MD&A announcement**

I next report the empirical results. My first model contains only the financial data variables that have been traditionally part of the every stock in SET50—the daily stock market return in the mean equation and its percentage change of trading volume in the variance equation. As expected, I find that this matters (see Table 12). Every stock in SET50 has its daily return move correlated to the overall stock market return, and larger change in trading volume tend to raise volatility in stock price movement.

Table 12 Tests of the Association between firms that announce MD&A and otherwise

	SET50 with MD&A report					SET50 without MD&A report				
	ADVANC	AOT	BCP	BDMS	BH	BA	BANPU	BBL	BLA	CBG
Mean Equation										
Market return	0.831 ***	0.848 ***	0.884 ***	0.639 ***	0.689 ***	0.041 ***	0.915 ***	0.773 ***	0.714 ***	0.197 ***
	(0.025)	(0.026)	(0.029)	(0.022)	(0.026)	(0.016)	(0.030)	(0.023)	(0.033)	(0.024)
Constant	0.000	0.000	-0.001 **	-0.001 *	0.000	-0.001 ***	-0.002 ***	0.000	-0.002 ***	-0.002 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Variance Equation										
Ch% volume	1.268 ***	0.927 ***	0.798 ***	1.033 ***	0.910 ***	0.818 ***	1.234 ***	0.888 ***	0.808 ***	1.184 ***
	(0.036)	(0.037)	(0.044)	(0.034)	(0.025)	(0.020)	(0.029)	(0.058)	(0.036)	(0.032)
Constant	-2.232 ***	-2.777 ***	-2.849 ***	-2.878 ***	-3.002 ***	-1.627 ***	-2.517 ***	-3.003 ***	-3.400 ***	-3.064 ***
	(0.219)	(0.363)	(0.447)	(0.337)	(0.399)	(0.140)	(0.220)	(0.575)	(0.415)	(0.161)
EGARCH terms										
1	0.752 ***	0.679 ***	0.670 ***	0.676 ***	0.648 ***	0.825 ***	0.706 ***	0.674 ***	0.600 ***	0.627 ***
	(0.025)	(0.043)	(0.053)	(0.039)	(0.048)	(0.016)	(0.027)	(0.063)	(0.050)	(0.020)
Observation	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,881
Log likelihood	5,600	5,386	5,379	5,588	5,303	6,105	5,188	5,978	5,241	5,215

Notes. This table presents coefficient estimates and z-statistic for baseline model. The z-statistics are constructed using EGARCH (1,1). Numbers in bracket are standard errors. */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively.

- **Tone and market volatility**

In the next step, I expand our model to include the tone and word_count variables in the variance equation. Results are reported in Table 13. The first noticeable result is that by controlling for the quantity of information, the estimated coefficients for the stock market return and percentage change of trading volume remain statistically significant. 6 from 25 listed firms, which disclose MD&A in English in any consecutive year, are responsive to the tone of the press releases. Only 3 stocks that are, on balance, optimistic (pessimistic), i.e., have a positive (negative) value for the tone variable tend to lower (raise) stock volatilities. While all of the log likelihood values seem to be improved by adding textual data into the model, the effect is economically small.

Table 13 The effect of tone in MD&A reports on stock price volatility

	ADVANC	BH	BTS	EGCO	INTUCH	PTTGC	SCB	SCC	TMB
Mean Equation									
Market return	0.832 ***	0.691 ***	0.623 ***	0.311 ***	0.766 ***	0.964 ***	1.089 ***	0.740 ***	0.886 ***
	(0.025)	(0.026)	(0.025)	(0.025)	(0.021)	(0.028)	(0.026)	(0.021)	(0.030)
Constant	0.000	0.000	0.000	0.000	0.000	-0.001 ***	0.000 *	0.000 *	-0.001 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Variance Equation									
Ch% volume	1.270 ***	0.908 ***	0.789 ***	0.669 ***	1.059 ***	1.025 ***	0.863 ***	0.961 ***	0.828 ***
	(0.036)	(0.025)	(0.048)	(0.040)	(0.026)	(0.060)	(0.055)	(0.053)	(0.036)
tone	-176.810 **	107.752 **	287.857 *	113.116 **	243.424 *	-152.724	-361.130 *	120.138	116.235
	(94.570)	(46.259)	(149.142)	(56.777)	(146.924)	(168.441)	(213.610)	(135.152)	(113.764)
word_count	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001 **	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	-2.294 ***	-3.025 ***	-2.331 ***	-2.945 ***	-2.047 ***	-3.164 ***	-2.621 ***	-2.760 ***	-2.619 ***
	(0.221)	(0.400)	(0.394)	(0.515)	(0.257)	(0.430)	(0.546)	(0.366)	(0.368)
EGARCH terms									
1	0.745 ***	0.646 ***	0.742 ***	0.673 ***	0.780 ***	0.633 ***	0.711 ***	0.704 ***	0.698 ***
	(0.026)	(0.048)	(0.045)	(0.057)	(0.029)	(0.051)	(0.061)	(0.040)	(0.044)
Observation	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880
Log likelihood	5,601	5,306	5,753	5,811	5,871	5,382	5,805	6,033	5,380

Notes. This table presents coefficient estimates and z-statistic of equation (5) and (6). The z-statistics are constructed using EGARCH (1,1). Numbers in bracket are standard errors. */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively.

- **Test of positive and negative sentiment**

Table 14 and 15 report results when I control for different components of the entire MD&A corpus. As mentioned above, I run separate regressions for each positive and negative sentiment variables. When differentiating between positive and negative sentiments, it is apparent that the positive words matters for stock price volatilities. Positive sentiments have a larger coefficient, and more significant level compare to negative sentiments. There are 6 listed firms that show statistically significant on positive sentiment while only 1 listed firms that show statistically significant on negative sentiments. Moreover, the EGARCH model that include positive sentiments yield result of log likelihood higher than model which include negative sentiments and tones. However, the relationship between positive sentiments and stock price volatilities are still inconclusive. 50% of listed firms that show statistically significant on positive sentiments have negative sign of coefficient referring that positive words in financial disclosure can reduce stock price volatilities while the other 50% listed firms demonstrate opposite explanation.

Table 14 The effect of positive tone of MD&A reports on stock price volatility

	ADVANC	BH	BTS	EGCO	INTUCH	PTTGC	SCB	SCC	TMB
Mean Equation									
Market return	0.834 ***	0.691 ***	0.624 ***	0.311 ***	0.764 ***	0.963 ***	1.089 ***	0.740 ***	0.902 ***
	(0.025)	(0.026)	(0.025)	(0.025)	(0.021)	(0.028)	(0.026)	(0.021)	(0.030)
Constant	0.000	0.000	0.000	0.000	0.000	-0.001 ***	0.000 *	0.000 *	-0.001 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Variance Equation									
Ch% volume	1.269 ***	0.906 ***	0.782 ***	0.664 ***	1.055 ***	1.029 ***	0.862 ***	0.960 ***	0.828 ***
	(0.036)	(0.025)	(0.049)	(0.040)	(0.026)	(0.060)	(0.055)	(0.053)	(0.036)
positive	-4.693 ***	1.488 **	2.711 **	-1.199	-0.708	-6.516 *	-2.511 **	0.860	2.625 *
	(1.525)	(0.587)	(1.373)	(1.666)	(1.811)	(3.658)	(1.253)	(2.872)	(1.531)
word_count	0.001 **	-0.001 **	-0.001 **	0.000	0.000	0.001	0.000 **	0.000	0.000
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Constant	-2.316 ***	-3.029 ***	-2.345 ***	-3.124 ***	-2.031 ***	-3.295 ***	-2.506 ***	-2.711 ***	-2.619 ***
	(0.218)	(0.401)	(0.398)	(0.531)	(0.258)	(0.429)	(0.541)	(0.363)	(0.363)
EGARCH terms									
1	0.742 ***	0.645 ***	0.740 ***	0.653 ***	0.782 ***	0.618 ***	0.723 ***	0.709 ***	0.698 ***
	(0.025)	(0.048)	(0.045)	(0.059)	(0.029)	(0.051)	(0.060)	(0.039)	(0.043)
Observation	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880
Log likelihood	5,605	5,306	5,753	5,810	5,870	5,383	5,805	6,032	5,381

Notes. This table presents coefficient estimates and z-statistic of equation (5) and (6). The z-statistics are constructed using EGARCH (1,1). Numbers in bracket are standard errors. */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively.

Table 15 The effect of negative tone in MD&A reports on stock price volatility

	ADVANC	BH	BTS	EGCO	INTUCH	PTTGC	SCB	SCC	TMB
Mean Equation									
Market return	0.832 ***	0.690 ***	0.621 ***	0.310 ***	0.764 ***	0.966 ***	1.090 ***	0.739 ***	0.888 ***
	(0.025)	(0.026)	(0.025)	(0.025)	(0.021)	(0.028)	(0.026)	(0.021)	(0.030)
Constant	0.000	0.000	0.000	0.000	0.000	-0.001 ***	0.000 *	0.000 *	-0.001 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Variance Equation									
Ch% volume	1.268 ***	0.909 ***	0.798 ***	0.667 ***	1.058 ***	1.021 ***	0.861 ***	0.959 ***	0.829 ***
	(0.036)	(0.025)	(0.048)	(0.040)	(0.026)	(0.060)	(0.055)	(0.053)	(0.036)
negative	-1.587	-1.680	-0.472	-1.861 **	-1.573	-0.039	-0.348	-2.961	3.087
	(2.102)	(1.586)	(2.589)	(0.889)	(0.996)	(2.364)	(1.257)	(1.959)	(2.994)
word_count	0.000	0.001	0.000	0.001 *	0.000	0.000	0.000	0.002 *	-0.001
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
Constant	-2.220 ***	-3.021 ***	-2.411 ***	-3.033 ***	-2.059 ***	-3.118 ***	-2.645 ***	-2.828 ***	-2.474 ***
	(0.221)	(0.399)	(0.400)	(0.517)	(0.257)	(0.430)	(0.557)	(0.370)	(0.341)
EGARCH terms									
1	0.753 ***	0.646 ***	0.732 ***	0.663 ***	0.779 ***	0.639 ***	0.708 ***	0.697 ***	0.715 ***
	(0.026)	(0.048)	(0.045)	(0.057)	(0.029)	(0.051)	(0.062)	(0.040)	(0.041)
Observation	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880
Log likelihood	5,600	5,304	5,749	5,812	5,871	5,381	5,803	6,033	5,381

Notes. This table presents coefficient estimates and z-statistic of equation (5) and (6). The z-statistics are constructed using GARCH (1,1). Numbers in bracket are standard errors. */**/*** denote statistical significance at the 10 percent, 5 percent, and 1percent levels, respectively, based on a two-tailed t-test.

Chapter 6
Summary and concluding remarks

Overview of the thesis

In this digital era, the advancement in technology with abundance of documents is kept in digital format. Due to plenty of space at low storage cost, managements tend to write financial disclosures longer, and provide more information for investors. I found that a number of Thai listed firms used MD&A report as a channel to communicate their firm's performance to business journalists, stock analysts and investors. A total of 47 listed companies launched MD&A report continuously from Q4/2011 - Q4/2018, although their profit might not change significantly from previous quarters. While many market participants have known and used quantitative data from MD&A reports frequently, my study can be considered the first to demonstrate modern text parsing methodology to extract the qualitative (textual) information from the big data set in the Thai stock market.

To summarize this study, I collected 1,421 MD&A samples from 49 companies. These documents contain about 6 million words in total, and then after I have preprocessed this qualitative data based on textual sentiment analysis, there are at least 3 million meaningful words left. I applied text parsing package in Python to tokenize and count words that have positive and negative sentiment according to Loughran and McDonald dictionary for finance. I found that, because of the long sample period under study, the tone of the press releases is close to zero on average (only slightly negative). There is a wide variation over time, with the tone ranging from -0.77 (very negative) to +1.00 (very positive).

Then I use Latent Dirichlet Allocation (LDA) to measure the proportion of topics discussed in these documents. The finding is that on average, 23% of management discussion report about firm's financial performance, 21 % discuss more about firm's financial status, 13% management raise topic about such external factor as global economy, export and regulations. As for the remaining 43% of discussion, managements report specifically about their firm's specific industry. However, considering solely on topic in order to explain future firm's performance may not be enough. I decided to use algorithm in python to count various tones according to Loughran and McDonald word list, which was created to use specifically in earnings press release documents. It is evident that managements normally use tonal words in FP topic to express about sentiment, while using uncertainty words in EF and FS topic when the condition is unclear.

My prominent research questions are whether this textual sentiment has a significant impact to future firm's performance and how stock market reacts to this information. According to my first hypothesis, in testing whether there is information about future firm's performance contained in MD&A reports, I found that the coefficient on textual sentiment variable was positive and significant suggesting that higher values of optimistic tone could predict higher future performance, and there is certain information in MD&A language incremental to be captured by other variables. To understand more about how managements disseminate the information about future firm's performance in MD&A documents, I use panel regression with firm fixed effect. The result shows that managements discuss more proportion on financial performance topic with more positive net tone, when future ROA is increasing.

Moreover, they tend to use ambiguous language such as “may,” “assume,” “possibly” in financial status topic, when managements expect firms’ profitability showing an upward trend. In addition, my result shows that topic proportion & tone estimated by whole corpus LDA is superior to sectoral corpus, in terms of future performance prediction power. One explanation could be that LDA is one kind of machine learning process, as such the more I provide input data, the more machine can cluster topics more efficiently.

For the second hypothesis, in testing prediction about unexpected change in net optimistic language in MD&A report, it is positively associated with market return around the announcement date. Interestingly, I found that investors reacted to this kind of information asymmetrically. The effect of the unfavorable tone of financial disclosures on stock market price is more pronounced and has more predictive power than the favorable tone. In addition, from the investment perspective, this study shows evidence that constructing portfolios with long top quartile and short bottom quartile of stock ranked by changes in tone could gain abnormal return. Additionally, my study also investigates whether investors value the topics & tones in MD&A documents, which are released simultaneously with financial statement in each quarter. I find that investors place greater value on management tone in firms’ financial status topic and external factor topic rather than tone in financial performance since the quantitative information about firm performance is easier to access and interpret by looking at net profit figure. On top of that, investors do value tone in financial status and external factor topic, because it may indicate additional value to stock market return in the next three-day period.

Finally, for the third hypothesis, I collected 700 MD&A samples from 25 companies. I employed a novel EGARCH (1,1) model to depict the results of net tone extracted from financial disclosure reports which impact the individual stock price volatility. The result shows that among the 25% of listed firms in SET50 which announce MD&A reports, the coefficients of the net tone are significant. Furthermore, I also study the difference in negative and positive sentiment variables. I found that the coefficients of the negative sentiment are insignificant across SET50 returns volatilities. Surprisingly, the coefficients of positive sentiment in EGARCH (1,1) are statistically significant and show negative signs in some stock returns. Since the good news is prevalent in the market, the overall volatility of market returns decreases. However, the result is not that robust after comparing with the estimated results from EGARCH (1,1) of other listed firms that also announce MD&A reports.

Adding the Flesch Reading Ease Score and Flesch-Kincaid Grade Level measure of readability of the press releases, I replicate the finding by Jansen (2011) that more complex statements raise volatility in financial markets. However, the coefficients of these readability measures in EGARCH (1,1) are statistically insignificant across the SET50 returns volatilities. More remarkably, I found that quantity of information variables (i.e. number of word count and number of sentence count) are statistically significant in some cases, which is similar with Loughran, McDonald and Yun (2013) findings that larger 10-Ks are significantly associated with high return volatility. The evidence is even stronger when I chose positive sentiments in the model, implying that positive words have more impact over net tone and negative words in predicting stock return volatilities in case of large-cap stocks in Thailand.

Implication of the results

For investors, according to my samples, this study has confirmed that the managers' tone in MD&A disclosures contain information about the firms' future performance. Therefore, investors and other stakeholders in The Stock Exchange of Thailand can use it as a primary source of inside information about company profitability in order to help the investment decision making. In addition, the implication from the investment perspective is that this study showed evidences that constructing portfolios with long top quintile and short bottom quintile of stock ranked by changes in tone could gain abnormal returns. If investors believed that textual sentiment has impact on stock returns, they would long top quintile portfolio and short bottom quintile portfolio, and expected that the profit would be the spread of implementing this investment strategy. The quintile spreads of each day forward after the MD&A announcement date were shown in figure 15. It was demonstrated that the investors were right. By longing top quintile and shorting bottom quintile of stock ranked by changes in tone could gain positive spread since day one to day fourteen. This can be implied that the impact of textual sentiment in MD&A disclosures had only short life (around 10 days forward), the maximum spread was about 0.7% in day four after the disclosures were released.

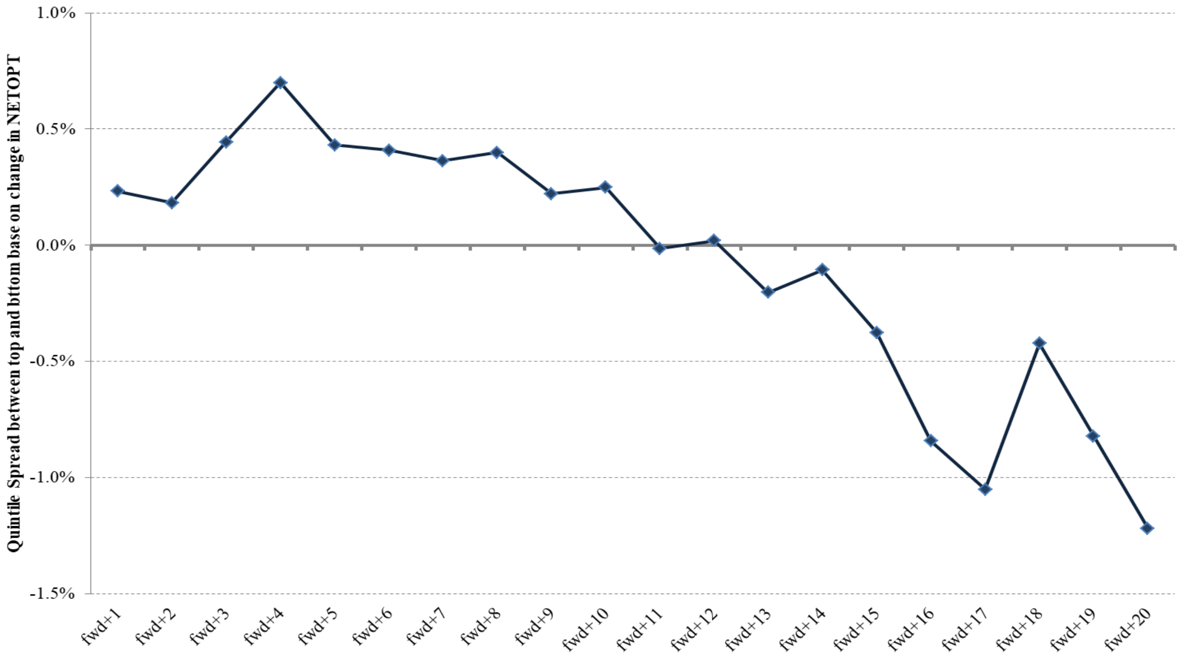

Figure 15 The quintile spreads after the MD&A announcement date.

Furthermore, I found that investors reacted to this kind of information asymmetrically. The effect of the unfavorable tone of financial disclosures on stock market price is more pronounced and has more predictive power than favorable tone. Figure 16 showed the four day buy-and-hold mean returns of each quintile portfolio, ranked by changes in tone. Although, the higher the value changes in net optimistic tone portfolios, the higher the average returns. The stated spread of 0.7% in buy-and-hold for four day came from 0.5% loses in bottom quintile and only 0.2% gain in top quintile. As a result, Investors seem to concern more about negative tone in management disclosure, and decided to sell stocks including in bottom quintile portfolio.

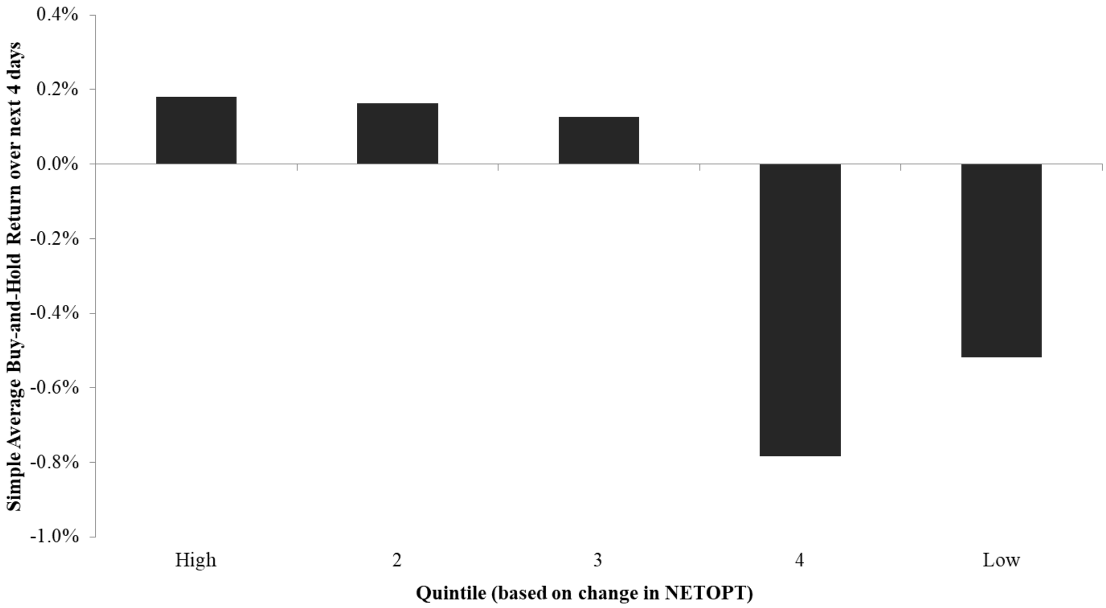

Figure 16 Four day buy-and-hold average returns of quintile portfolios

For researchers, my study can be considered among the first to demonstrate modern text parsing methodology to extract the qualitative (textual) information from the big data set in the Thai stock market. It indicated that measuring textual sentiment by using dictionary-based approach with financial dictionary from the US can be applied to English-Thai financial disclosure reports. It referred that Thai managements write their financial reports using as similar word list as those management in US. Additionally, a complicated probabilistic machine learning model like LDA can work well in measuring topic proportion in earning announcement in Thailand. I view that these two textual analysis tools can be a very good start in applying other tools to quantify words. More importantly, I have constructed new data sets such as textual sentiment and proportion of topic for other researchers to use and create more studies in related fields.

For management in listed companies, the study concludes that investors responded to tone in firm's financial status and external factor topic rather than firm's financial performance topic. Therefore, managements should not only provide explanation about firm's profitability, but also write about outlook of the economy, competition level among rivals, demand and supply of products or services and risks that shareholders should be concerned. By looking at readability measures, Thai financial disclosures are still very difficult to read, especially complicated sectors like healthcare and petrochemicals. Moreover, listed firms in SET50 that announce MD&A should provide precise text information in the reports, because the longer the document, the more difficult for investors to digest information, causing higher stock price volatilities.

For policy makers, information is always crucial in the stock market. Listed firms should therefore be encouraged to fully disclose their financial performance and any other information which will benefit debtors and shareholders, not only in terms of quantity, but also in terms of quality. Today the text parsing technology is developing to a certain extent, allowing regulators to monitor these financial disclosures more easily. To make sure that the capital market is fair, transparent and efficient, regulators should encourage firms to transform any information into digital format to take advantage of the advancement of machine readable technology.

Limitation of study

The knowledge in textual analysis of this study is limited for only English language while most financial disclosures in The Stock Exchange of Thailand are still in Thai. However, when comparing the contents between Thai and English MD&A reports, I found that most of them are identical. By choosing only MD&A in English version makes the samples in this study are comparatively small. Only 47 from the total of 700 listed companies passed my criteria. Most of them are large-cap stocks, which contribute about 50% of total market capitalization in The Stock Exchange of Thailand. Although dictionary-based approach and topic modeling are processed efficiently in this sample, meanwhile some of the text parsing techniques like readability is not. One explanation is that many Thai-English technical words are included in MD&A. Lastly, due to small sample size, it was not practical to group stocks into factor base model according to previous literature in stock market return prediction.

Area for future research

For future research, new qualitative information sources and other textual analysis approaches including statistic model that have not yet been widely used are desirable for future studies. The main objective in develops the textual analysis is to create the new approach that can measure tone and topic of documents more efficiently. Since many firms and investors rely more on the result from textual sentiment analysis and topic modeling. For example, third party valuation judge the companies level in responsible for environmental, social and governance (ESG) by proportion of topic that firms disclose in annual report, or listed firms that invest more in product line that gain popularity, measured by textual sentiment, in online social media. Meanwhile, it is interesting to analyze qualitative information in other languages beside of English, as different markets may display differing cultural and other behavioral patterns. For textual analysis in Thai language, the most challenge is still the word tokenization, since Thai language sentence has no space between words and we don’t use full stop at the end of each sentence. Finally, I hope the essence of this research will be answers to numerous questions and a useful tool for investors and stakeholders of the capital markets locally and globally, while further discovery beyond this will benefit the capital market as a whole.

REFERENCES

Blei, David M, Andrew Y Ng, and Michael I %J Journal of machine Learning research Jordan. "Latent Dirichlet Allocation." 3, no. Jan (2003): 993-1022.

Bodnaruk, Andriy, Tim Loughran, Bill %J Journal of Financial McDonald, and Quantitative Analysis. "Using 10-K Text to Gauge Financial Constraints." (2015): 623-46.

Das, Sanjiv R, and Mike Y %J Management science Chen. "Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web." 53, no. 9 (2007): 1375-88.

Davis, Angela K, Weili Ge, Dawn Matsumoto, and Jenny Li %J Review of Accounting Studies Zhang. "The Effect of Manager-Specific Optimism on the Tone of Earnings Conference Calls." 20, no. 2 (2015): 639-73.

Davis, Angela K, J Piger, and Lisa M Sedor. "Beyond the Numbers: An Analysis of Optimistic and Pessimistic Language in Earnings Press Releases. Federal Reserve Bank of St." Louis Working Paper No. 2006-005A, 2006.

Doran, James S, David R Peterson, S McKay %J The Journal of Real Estate Finance Price, and Economics. "Earnings Conference Call Content and Stock Price: The Case of Reits." 45, no. 2 (2012): 402-34.

Durnev, Art, and Claudine %J Journal of Accounting Research Mangen. "Corporate Investments: Learning from Restatements." 47, no. 3 (2009): 679-720.

Ehrmann, Michael, and Jonathan %J Journal of Monetary Economics Talmi. "Starting from a Blank Page? Semantic Similarity in Central Bank Communication and Market Volatility." 111 (2020): 48-62.

Engelberg, Joseph. "Costly Information Processing: Evidence from Earnings Announcements." Paper presented at the AFA 2009 San Francisco meetings paper, 2008.

Engle, Robert F %J Econometrica: Journal of the Econometric Society. "Autoregressive Conditional

Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." (1982): 987-1007.

Fama, Eugene F %J The journal of Finance. "Efficient Capital Markets: A Review of Theory and Empirical Work." 25, no. 2 (1970): 383-417.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin %J Review of Accounting Studies Segal. "Management's Tone Change, Post Earnings Announcement Drift and Accruals." 15, no. 4 (2010): 915-53.

Ferris, Stephen P, Qing Hao, and Min-Yu %J Review of Finance Liao. "The Effect of Issuer Conservatism on Ipo Pricing and Performance." 17, no. 3 (2013): 993-1027.

Griffiths, Thomas L, and Mark %J Proceedings of the National academy of Sciences Steyvers. "Finding Scientific Topics." 101, no. suppl 1 (2004): 5228-35.

Hansen, Stephen, Michael McMahon, and Andrea Prat. "The Dilemma of Central Bank Transparency." Centre for Economic Performance, LSE, 2015.

Henry, Elaine %J The Journal of Business Communication. "Are Investors Influenced by How Earnings Press Releases Are Written?". 45, no. 4 (2008): 363-407.

Huang, Allen H, Reuven Lehavy, Amy Y Zang, and Rong %J Management Science Zheng. "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach." 64, no. 6 (2018): 2833-55.

Huang, Xuan, Siew Hong Teoh, and Yinglei %J The Accounting Review Zhang. "Tone Management." 89, no. 3 (2014): 1083-113.

Jegadeesh, Narasimhan, and Di %J Journal of financial economics Wu. "Word Power: A New Approach for Content Analysis." 110, no. 3 (2013): 712-29.

Kearney, Colm, and Sha %J International Review of Financial Analysis Liu. "Textual Sentiment in Finance: A Survey of Methods and Models." 33 (2014): 171-85.

Li, Feng %J Journal of Accounting, and economics. "Annual Report Readability, Current Earnings, and Earnings Persistence." 45, no. 2-3 (2008): 221-47.

Li, Feng %J Journal of accounting literature. "Survey of the Literature." 29 (2010): 143-65.

Loughran, Tim, and Bill %J Journal of Accounting Research McDonald. "Textual Analysis in Accounting and Finance: A Survey." 54, no. 4 (2016): 1187-230.

Loughran, Tim, and Bill %J The Journal of Finance McDonald. "Measuring Readability in Financial Disclosures." 69, no. 4 (2014): 1643-71.

———. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." 66, no. 1 (2011): 35-65.

Loughran, Tim, Bill McDonald, and Hayong %J Journal of Business Ethics Yun. "A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports." 89, no. 1 (2009): 39-49.

MacKinlay, A Craig %J Journal of economic literature. "Event Studies in Economics and Finance." 35, no. 1 (1997): 13-39.

Price, S McKay, James S Doran, David R Peterson, Barbara A %J Journal of Banking Bliss, and Finance. "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone." 36, no. 4 (2012): 992-1011.

Szarowska, Irena %J International Journal of Monetary Economics, and Finance. "Effect of Macroeconomic Determinants on Non-Performing Loans in Central and Eastern European Countries." 11, no. 1 (2018): 20-35.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus %J The Journal of Finance Macskassy. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." 63, no. 3 (2008): 1437-67.

Wu, Di %J Unpublished working paper. "Shock Spillover and Financial Response in Supply Chain

Networks: Evidence from Firm-Level Data." (2016).

VITA

NAME	Siriyos Chuthanondha
DATE OF BIRTH	27 Jan 1983
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	The Stock Exchange of Thailand
HOME ADDRESS	24/10 Inthamara 51 Dindeang Dindeang Bangkok 10400
PUBLICATION	Int. J. of Monetary Economics and Finance