Scholarship Project Paper 2019

Capturing Investor Sentiment from Big Data: The Effects of Online Social Media on SET50 Index

Associate Professor Nongnuch Tantisantiwong, Nottingham Trent University, UK
Dr. Kulabutr Komenkul, Rangsit University, Thailand
Dr. Charika Channuntapipat, University of Birmingham, UK
Dr. Watthanasak Jeamwatthanachai, University of Southampton, UK

20 April 2020

Abstract

This research aims to introduce a market sentiment index which can be used as a leading indicator for the Stock Exchange of Thailand (SET). This new index is constructed from Big Data extracted from online social media. The data used in this project are the daily firm-level and market data, market capitalizations of firms and the SET, the total trading value in the SET and the values of stocks bought and sold by investor types, the data on S&P global index, the implied volatility index of S&P500 (known as VIX) and the exchange rate of Thai Baht. Meanwhile, tonal words regarding 50 companies listed in the SET50 index are extracted from selected online/social media using search engines and an application programming interface (API). Note that 50 firms included in the analyses and index construction change every six months following the announcement of the SET. The daily data will be collected for the period 2015 - 2018. Machine Learning algorithms are employed to conduct a bag-of-word analysis which will count the number of positive, negative and neutral tonal words for 50 firms in a set of articles over a 4-year period. The findings show that the introduced sentiment index has significant relationships with the changes in SET50 index. The sentiment index can signal changes in SET50 index from day t to day t+3. The analysis results indicate that this sentiment index can act as a leading indicator of the SET50 index and thus the SET index.

JEL Classification: G10, G11, G14
Keywords: Stock price return, Prediction, Sentiment Analysis, Text Mining
E-Mail Address: kbkomenkul@gmail.com

Content

Figure

Executive Summary

This project aims to introduce a market sentiment index which can be used as a leading indicator for the Stock Exchange of Thailand (SET). This new index is constructed from Big Data extracted from online social media. We introduce a method to construct an index which can conclude whether investors are overall pessimistic or optimistic toward listed firms based on investors' perspective expressed in Thai online social media. Here, we also show that there exists a significant positive relationship between the change in SET50 index and our sentiment index, proving that our sentiment index can reflect market sentiment and investor perception toward the performance of stocks listed in the SET and can be a good leading indicator for the stock market. This research project is among the first to construct a conclusive market sentiment index that can act as a leading indicator of the stock market and analyze the relationship between the stock market performance and tonal information from online social media (i.e. sentiment) by using textual analysis in Thai language.

The data used in this project are obtained from various sources. The daily firm-level and market data such as market capitalizations of firms and the SET, the total trading value in the SET and the values of stocks bought and sold by investor types are retrieved from the database of Stock Exchange of Thailand and SETSMART. The data on S&P global index, the implied volatility index of S&P500 (known as VIX) and the exchange rate of Thai Baht are obtained from Datastream. Meanwhile, tonal words regarding 50 companies listed in the SET50 index are extracted from selected online/social media using search engines and an application programming interface (API). Note that 50 firms included in the analyses and index construction change every six months following the announcement of the SET. The daily data will be collected for the period 2015 - 2018. A set of 10 Thai online social media platforms are selected for this project, based on their popularity (e.g. visibility in online community, and number of LIKEs or subscribers) and relevance (e.g. specializing in providing financial and investment decision).

The research team develops a dictionary of Thai tonal words appearing in online social media which show public emotion and perception toward companies listed in the SET50 index. The tonal words are used in combination with identified directional and contextual words to capture the sentiment on the SET50 companies. Machine Learning algorithms are employed to conduct a bag-of-word analysis which will count the number of positive, negative and neutral tonal words for 50 firms in a set of articles over a 4-year period. These numbers are used to create a new sentiment index. Next, a multiple regression is used to test whether the SET50 index is indeed associated with this new sentiment index. The findings show that the introduced sentiment index has significant relationships with the changes in SET50 index. The sentiment index can signal changes in SET50 index from day t to day t+3. The analysis results indicate that this sentiment index can act as a leading indicator of the SET50 index and thus the SET index.

The outcome and deliverables of the project therefore include this project report which contains the finding on the relationship between the SET50 index and our new sentiment index; a sentiment index which can reflect the investors' mood and be a new leading indicator of the stock market performance adding to the current set of leading indicators that have been used at the SET and other organizations; and a new approach to construct a composite sentiment index which can eliminate the conflict of signals provided by different sources of information.

These findings and outcomes will not only benefit local investors, but also foreign investors as it will potentially provide a simplified leading indicator derived from sentiment from Thai online social media sources.

Chapter 1 Introduction

New information is believed to affect the movement of the stock market. A plethora of studies have examined the stock market reaction to companies' announcements. However, the booming of online and social media has recently shortened the time that new information becomes available to investors. Online social media also facilitate the availability of opinion-led information in addition to factual information from companies' announcement; the public can use such information to form their investment decisions. Thus, the information from these online opinion-led platforms (i.e. social media, personal blogs, and forums), which embed investors' sentiment, could also have an impact on the movement of stock prices (Dergiades, 2012; Verma and Verma, 2007).

The majority of research exploring the relationship between market reaction and tonal information from online and social media (i.e. sentiment) has been done in the context of English language (see e.g. Kim and Kim, 2014; Sun et al., 2016) and also of Chinese language (see e.g. Day and Lee, 2016; Guo et al., 2017). However, no studies have been done using Thai social media.

Unlike existing studies, the project aims to propose a new media and investment sentiment index which can be used as a *leading indicator* for the Stock Exchange of Thailand (SET). Big Data from online and social media sources will be used to develop the indicator. In particular, the information that we are interested in is investors' sentiment information appearing in online and social media in Thai language.

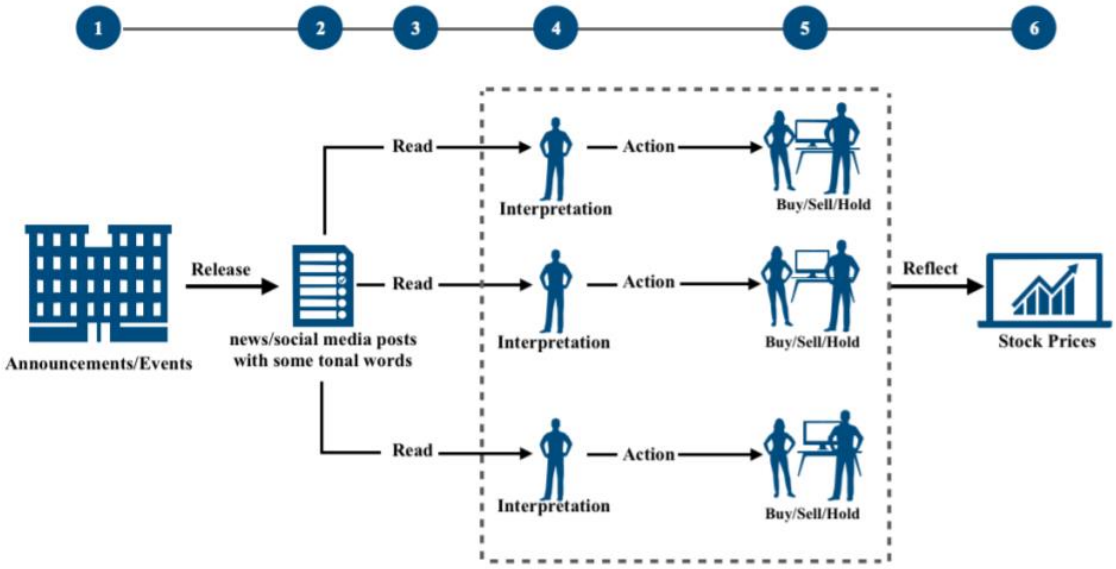
1.1 Research motivations and context

This research is motivated by the investors' reliance on opinion-led information from online and social media; the effects of investors' reaction to information on stock prices; the availability of Big Data; and the development of Big Data Analytics. This leads us to explore the potential leading effect of Big Data on the Thai stock market. According to the statistics from Globalstats Statcounter (2018), the top four social media used by Thai people as at November 2018 are

Facebook (71.72%), Twitter (16.98%)[1], Google+ (5.77%), YouTube (3.06%). From Facebook statistics, 6 out of top 10 Thai Facebook pages that have the largest numbers of fans are Thai media. On top of this, the data from Global Web Index shows that Thailand is ranked as the world leader for time spent on the internet and mobile internet usage with an average of 9 hours 38 minutes per day (We are social, 2018). This new lifestyle of Thai people urges the necessity to develop an up-to-date and dynamic leading indicator from online sources of information.

Individual investor may interpret information they receive differently, depending on their experience, knowledge, mood, or other influencing factors. (This is supported by behavioral finance and investor sentiment theories that investors' behavior can be shaped by whether they feel optimistic (bullish) or pessimistic (bearish) about future market returns. Their reaction to information may then impact the market (see Figure 1).

Figure 1: The impact of information available on stock prices



This project only focuses on information from online social media in which news and articles are entirely written in Thai and, to some extent, include information about firms or investors' views and recommendation on trading strategies. The online social media used in this project provide local news, comments and opinions in Thai and thus extracted tonal words are valuable information that can indicate market sentiment and may influence the market movement. According to information from the SET, in 2018 (yearly cumulative number from 1 January 2018

[1] While the proportion of Facebook users in Thailand has reduced from 94.45% in November 2017 to 71.72% in November 2018 (of the total number of Thai social media users), the proportion of Twitter users has increased from 1.18% to 16.98% on the same period.

to 25 December 2018), local individual investors had 4% more buying transactions and 6% more selling transaction than foreign investors did (The Stock Exchange of Thailand, 2018)[2]. It is, therefore, important to capture market sentiment from the local information sources written in the local language.

To our knowledge, this project is the first study to construct a composite sentiment index from Thai online social media. In order to prove that this index can act as a leading indicator of the stock market performance, we analyze the relationship between investors' sentiment exposed in online social media, composed in Thai, and stock market index changes. The result from this analysis would add to existing studies which provided evidence on investors' sentiment changes that appear to lead stock price changes especially for stocks heavily traded by individuals (Yang et al., 2017; Verma and Verma, 2007), and traded during the extreme market movement (Baek, 2016).

Our new sentiment index will be useful for not only local investors but also foreign investors because it converts sentiment embedded in online social media which is in such language that is not understandable for foreign investors to be a simplified leading indicator.

Given that Thai stock market indexes are leading indicators for Thai economic outlook, investors and policymakers can also use this proposed index as a leading indicator to monitor and forecast Thai economic growth as well as the prediction of excessive speculation of stock trading (Yang and Zhang 2014).

1.2 Research objectives and research questions

This project aims to introduce a sentiment index which can be used as a leading indicator for the SET, not to neither predict the movement of the SET or SET50 index nor find the best set of determinants for the SET or SET50 index changes. This new index can be used by the SET and any policymakers as a leading indicator to monitor and predict the movement of market and Thai economy, and by investors as information considered when making investment decision (Sayim and Rahman, 2015).

A Big Data analytic process is employed to extract tonal words from Thai online social media which will then be used to construct a composite sentiment index reflecting the, on average, pessimism/optimism toward listed firms expressed in Thai online and social media. In particular,

[2] Local individuals (Foreign investors) account for approximately 40% (36%) of the total value of buying transactions and 32% (38%) of the total value of selling transactions.

tonal words are extracted from articles relating to the firm constituents of the SET50 index. Tonal words could reflect investors' sentiment to the events of listed firms, and hence the index created from the bag-of-words analysis would, to some extent, also represent market sentiment.

1.3 The scope of study

Due to time and budget limitation, this project will focus only on the information relating to SET50 companies between 2015 and 2018. These 50 firms have the largest market capitalization in the SET, representing 67.73% of the total market capitalization of the SET in 2018 (72.50% in 2014), and their stocks are highly liquid. As a result, the SET50 index's changes are highly correlated with changes in the SET50 index.

The 10 selected online and social media websites are used as information sources for this project. We do not directly use information from analysts' reports as immediate sources for the analysis as the information expressed is, to some extent, integrated within the social media. In addition, although a lot of investors rely on such reports for investment decisions, some analysts' reports are not publicly available.

The data employed in this project cover the period 2015-2018. There are some reasons for choosing this data period. First, during 2005 - 2014 the stock market index and stock prices may be affected by economic and political turmoil both in Thailand and in other countries such as the global imbalance in 2005-2006, the sub-prime crisis in 2006-2007, the global financial crisis in 2007-2009, the Euro debt crisis in 2009-2014 and political unrests in Thailand during 2005-2014. Second, there may also be some significant impacts of the Quantitative Easing policies which were employed in the US, UK, EU and Japan since 2009. Third, findings from Sun et al. (2016) and Chung et al. (2012) show that predictive power of investor sentiment is stronger for expansion states, not recession states. Last but not least, the 4-year sample period between 2015 and 2018 includes the period when social media and internet usage in Thailand increases[3].

1.4 Structure of this report

The remainder of this report is outlined as follows. The next section discusses relevant theories and reviews prior empirical studies. Section 3 describes data and methods employed to construct the sentiment index and to prove that the new sentiment index can act as a leading indicator for the Thai stock market. In Section 4, a plot and descriptive statistics of the new sentiment index

[3] The number of Thai internet users only account for 30% of the population in the mid of 2012, but social media penetration rate for Thailand is 74% in 2017.

are presented and the relationship between the SET50 index changes and the sentiment index is analyzed. Finally, Section 5 concludes.

Chapter 2 Extant Literature

The research question and objectives of this study are informed by three main areas of extant literature namely: behavioral finance (establishing finance and investment sense), linguistics (understanding the nature of language), and machine-learning (understanding computational modelling). The following literature review provides background theories and concepts to develop a fundamental understanding of this research project.

2.1 Behavioural finance, social media, and Big Data applications

Fama (1965) introduced a theory of efficient market hypothesis (EMH) that the market is not predictable. The author put forth the fundamental idea that it is virtually impossible to consistently “beat the stock market” by making investment returns that outperform the overall market average as reflected by major stock indexes such as the S&P 500 Index. According to Fama’s theory, while an investor might get lucky and buy a stock that brings huge short-term profits, over the long term they cannot achieve a return on investment that is substantially higher than the market average. Later, Fama (1970) revised the EMH and classified market efficiency into 3 levels (strong, semi-strong and weak). While the strong form of efficiency is about the inability of investors to beat the market using insiders’ information, the semi-strong and weak forms of efficiencies refer to the inability of investors to beat the market using any public available information and any past information, respectively. Recently, Urquhart and Hudson (2013) introduced a counter-theory to the EMH namely, Adaptive Market Hypothesis (AMH). They investigated the AMH in three markets (US, UK and Japan) by using very long run data. They found an evidence of returns going through periods of independence and dependence. Although the magnitude of dependence varies over time, nonlinear dependence is strong throughout. In addition, they suggested that the AMH provides a better description of the behavior of stock returns than the EMH. Many studies document that the weak or semi-strong form of inefficiency has recently disappeared in mature markets. However, some researchers find that several emerging markets are at least not semi-strong-form efficient.

Given that Big Data (tonal words) from online and social media are analyzed within this project, our analysis is much relevant to the above hypotheses. If the sentiment hidden in online

public posts has a significant impact to stock prices/indexes in the current or next period, then the finding indicates the existence of market inefficiency.

With the advent and popularity of free internet access, social media have become a part of nearly everyone's daily routine. "From a technical point of view, social media are web-based or mobile technologies necessary for operating of highly interactive platform where users create, modify and share user-generated content" (Bukovina, 2011, p. 71). Thus, anyone is virtually connected, all the time, whether they are texting each other on social media platforms like WhatsApp, Line, or WeChat; checking or sharing posts or news on Facebook or other online media; or posting updates on Twitter, and so on. Social media sites can be distinguished from earlier forms of media by their more dynamic, message-based, interactive, and socially networked nature. Regarding to Statista (2018), as of October 2018, there are over 2 billion users on Facebook, 803 million on China's Tencent QQ, 336 million on Twitter, 303 million on LinkedIn, 1 billion on Instagram, and 250 million active users on Pinterest.

Due to the fact that conversations, dialogues, and opinions are initiated and shared constantly via social media platforms, these platforms can be used to share/gather information that is relevant to stock market. Social media are unique in facilitating substantial firm-investor interactions. Some websites and blogs, although having one-way communication from financial analysts to investor, allow sharing content between investors, while the internet message boards that have been the subject of considerable prior research are many-to-many (investors only). These are the sources of the "Big Data".

Big Data obtained from social media (hereafter known as social media Big Data) refers to information, opinions, and activity of individuals, interactions among them or more precisely the complex behavior of a society. In behavioral finance field, society's behavior and its relation to capital markets are dominant parts of analysis in the field of behavioral finance. As a result, the behavioral finance framework serves as a main motivation for the employment of social media Big Data in stock markets. In addition to this, behavioral finance challenge the notions of efficient markets. Evidence from the social media, communication, and capital markets literature suggests that the amount and speed of information has dramatically increased, and that this increased information can impact the stock market movements (Joseph et al. 2011). Such media do not only provide information on the status of the market, but also have an impact on market sentiments based on the released news or information (Nassirtoussi et al., 2014).

Although Big Data provided by social media are not the only source of data adopted in the analysis of capital market, several recent studies (see e.g. Bollen et al., 2011; Davis et al., 2006; Joseph et al. 2011; Li, 2006; Tetlock et al.,2008; Zhang et al., 2011; Li et al., 2014) employ social media Big Data in their analysis. The analysis process normally involves transforming qualitative information on various social media into quantitative form, and analyzing them to find relationships between the information and the stock market sentiments. In other words, qualitative data from Twitter, Facebook, blogs and websites are quantified and used in a regression analysis.

2.2 Analytical approaches to capture investor sentiment

Sentiment analysis employed in finance research is developed from various sentiment dimension. Li et al. (2014) define the sentiment analysis as a text analysis, the use of natural language processing and computational linguistics to identify and extract subjective information in source materials.

Stock market prediction using social media and textual web data has attracted much attention from academia as well as business. Li (2006) and Davis et al. (2006) studied the tone of qualitative information using objective word counts from corporate annual reports and earnings press releases, respectively. Li (2006) showed that the words “risk” and “uncertain” in firms’ annual reports predict low annual earnings and stock returns. Zhang et al. (2011) investigated the stock market indicators through Twitter. They collected the twitter feeds for 6 months and received a randomized sub sample of about one hundredth of the entire tweets, as the total of volume was about 2.5 million tweets per day. They analyzed the positive and negative mood of the masses on the twitter comparing to the stock market indices (on day t+1) such as Dow Jones, S&P500, and NASDAQ. They found that emotional tweet percentage significantly negatively related to the stock market indices, but it is positively related to VIX (the Chicago Board Options Exchange Volatility Index). Similarly, Bollen et al. (2011) investigate the relationship between measurements of Twitter moods and the value of the Dow Jones Industrial Average index (DJIA). Their results show that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions. Their model has an accuracy of 86.7% in prediction of daily DJIA movements.

Bag-of-words based approaches

Bag-of-words based approaches model news articles by vector space model which translates each news piece into a vector of word statistical measurement such as the number of

occurrences. The Bag-of-words scheme represents all words appearing in news as a document-term matrix. For instance, Tetlock et al. (2008) demonstrate a row could be the 1/8/99 Microsoft story above, and columns could be the terms “alleged,” “abuse,” “worse,” “happy,” and “neutral.” The matrix elements are designed to capture the information value of each word in each news story, which could be the relative frequencies of the 5 words within the 29-word excerpt: [1/29, 1/29, 1/29, 0/29, 0/29]. The challenge in text analysis is to translate this term-document matrix into a meaningful conceptual representation of the story, such as the degree to which the story conveys positive or negative information.

Instead of just measuring word frequency, each word is decomposed and represented by a vector of sentiment features. For example, word “accelerate” can be represented by strong and active by using Harvard IV-4 psychological dictionary. Each document can be represented by a vector of sentiment values by summing up the sentiment vectors of each word in the document.

Sentiment Analysis in Finance Research

The sentiment analysis of news articles and their impact on stock price returns have been studied in finance domain. For example, Tetlock et al. (2008) found that a simple quantitative measure of language can predict individual firms’ accounting earnings and stock returns. Their analysis involved only a fraction of negative words in Dow Jones News Service (DJNS) and Wall Street Journal (WSJ) stories relating to S&P 500 firms from 1980 to 2004. They quantified the language used in financial news (earning announcements, mergers or analysts’ recommendations) in an effort to predict firms’ accounting earnings and stock returns. They found that (i) the fraction of negative words in firm-specific news stories forecasts low firm earnings, (ii) firms’ stock prices underreact to the information embedded in negative words and (iii) the earnings and return predictability from negative words is largest for the stories that focus on fundamentals. More recently, Li et al. (2014) have investigated the impacts of a major financial news vendor in Hong Kong on stock return by using sentiment analysis.

2.3 Sentiment information from different sources and its effect on stock price movement

While some studies might find that sentiment data can generally have an impact on stock price movement, others found that sentiment information has more impact on stock price movement under certain conditions. Sul et al. (2017) use Twitter as a social media platform to capture the sentiment and document that “Tweets” from the social media platform can have an impact on stock price movement. More importantly the more those tweets are retweeted, the

faster the impact becomes observable on stock price movements. This can also be interpreted that the retweeted tweets can be considered as important information and catch the attention of Twitter users. This is aligned with the findings from Guo et al. (2017) that the sentiment information can be used to predict the stock price only when the stock has high investor attention.

This claim is also supplemented by the study of Bajo and Raimondo (2017), which explored media sentiment and IPO pricing. Using information from 2,800 US IPOs and sentiment information from over 27,000 articles, they found that tonal information, timing, and media reputation are important factors affecting the stock price movements (Bajo and Raimondo, 2017). Similar to the finding from Sul et al. (2017), Bajo and Raimondo's (2017) study reflects that media tones are important only if coupled with investor attention.

Yu et al. (2013) used daily media content from different social media outlets, including blogs, forums, and Twitters to observe stock price movements of 824 traded companies' 52,746 messages. They concluded that information from both social media and conventional media have a strong interaction effect on stock performance. Moreover, the findings also show that information from social media outlets has a stronger relationship with the stock performance than that from conventional media; and that different types of social media do have different impact on stock price movement (Yu et al., 2013). Although the findings indicate that types of social media do have different impact on stock price performance, it was conclusive that stock returns are associated with information from social media (Yu et al., 2013).

Day and Lee (2016) classified media sources, not by platforms, but by media specialty (i.e. Finance and Economic focus or not); and explore the relationship between different classification and the impact of the sentiment information on stock performance. They explored news articles from four news providers and capture the sentiment on financial performance from those news. They found that sentiment information from various financial resources have significantly different effects on investors' investment decisions. Financial news from media domains that are specific to Finance and Economics have more impact on stock price performance (Day and Lee, 2016). This is linked to the previous claim stating that sentiment information would have more impact if coupled with investors' attention, which can come from the popularity of the news/incidents relating to firms, and coverage and popularity of the media sources. This is one of the reasons why in this study we aim to choose ten sources of media based on their reputation.

Chapter 3 Methodology

In order to construct a new leading indicator for the SET, Big Data analytics is employed here. More specifically, the project uses the daily data on market capitalization and tonal words from online media, which intrinsically reflect the mood of investors, for firms listed on the SET50 index. The SET50 index is a market capitalization - weighted price index of 50 largest and most liquid stocks. Below is the description of the sentiment index construction process and the test we employ to prove that the index can act as a leading indicator of the market.

3.1 Data collection

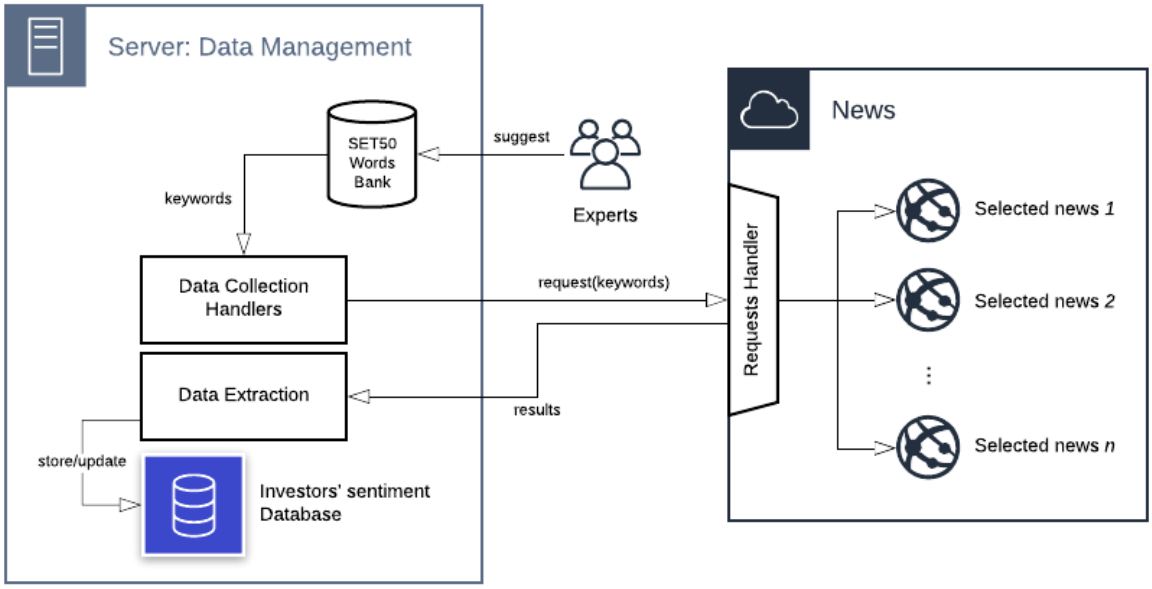
The data employed in this project are obtained from various sources. Daily data on market capitalization for each of SET50 constituents and Thai equity market-related information (e.g. constituents of the SET50, the SET50 index, stock trading statistics) are obtained from the database available at Stock Exchange of Thailand. Other variables such as S&P Global index and VIX are collected from the Datastream database.

Meanwhile, tonal words are extracted from news and posts regarding 50 companies listed in the SET50 index which were published in selected online/social media using search engines and an applicationn programming interface (API). Figure 2 shows the workflow for data collection.

Both of quantitative and qualitative data were collected for the period 2015—2018. The period during 2005 – 2014 is excluded from our study for the following reasons. The stock market index and stock prices may be affected by economic and political turmoil both in Thailand and in other countries during 2005-2014. These events include the global imbalance in 2005-2006, the sub-prime crisis in 2006-2007, the global financial crisis in 2007-2009, the Euro debt crisis in 2009-2014, political unrests in Thailand during 2005-2014, and the national election in 2019. There may also be some significant impacts of frequent changes in Quantitative Easing policies, which were employed in the US, UK, EU and Japan, especially during 2009-2014.

Moreover, findings from Sun et al. (2016) and Chung et al. (2012) show that predictive power of investor sentiment is stronger for expansion states, not recession states. Further, the 4-year sample period between 2015 and 2018 includes the period when social media and internet usage in Thailand increases.

Figure 2: Data Collection Workflow



Tonal words collected from 10 different sources listed below[4].

1) EfinanceThai: http://www.efinancethai.com/index.aspx
2) Investerest: http://www.investerest.co/
3) Kaohoon: https://www.kaohoon.com/
4) Khaosod: http://www.khaosod.co.th/
5) Longtunman: http://longtunman.com/
6) Matichon: http://matichon.co.th/
7) Manager Online: http://mgronline.com/
8) Post Today: http://posttoday.com/
9) Thairath: http://www.thairath.co.th/
10) Thunhoon: https://www.thunhoon.com/

[4] As mentioned earlier, the media sources were selected based on their popularity and relevance. Although the list is not exhaustive, it provides a sound starting point for the frequently used and relevant sites.

Both of quantitative and qualitative data were collected for the period 2015—2018. The period during 2005 – 2014 is excluded from our study for the following reasons. The stock market index and stock prices may be affected by economic and political turmoil both in Thailand and in other countries during 2005-2014. These events include the global imbalance in 2005-2006, the sub-prime crisis in 2006-2007, the global financial crisis in 2007-2009, the Euro debt crisis in 2009-2014, political unrests in Thailand during 2005-2014, and the national election in 2019. There may also be some significant impacts of frequent changes in Quantitative Easing policies, which were employed in the US, UK, EU and Japan, especially during 2009-2014.

Moreover, findings from Sun et al. (2016) and Chung et al. (2012) show that predictive power of investor sentiment is stronger for expansion states, not recession states. Further, the 4-year sample period between 2015 and 2018 includes the period when social media and internet usage in Thailand increases.

3.2 Objective variables

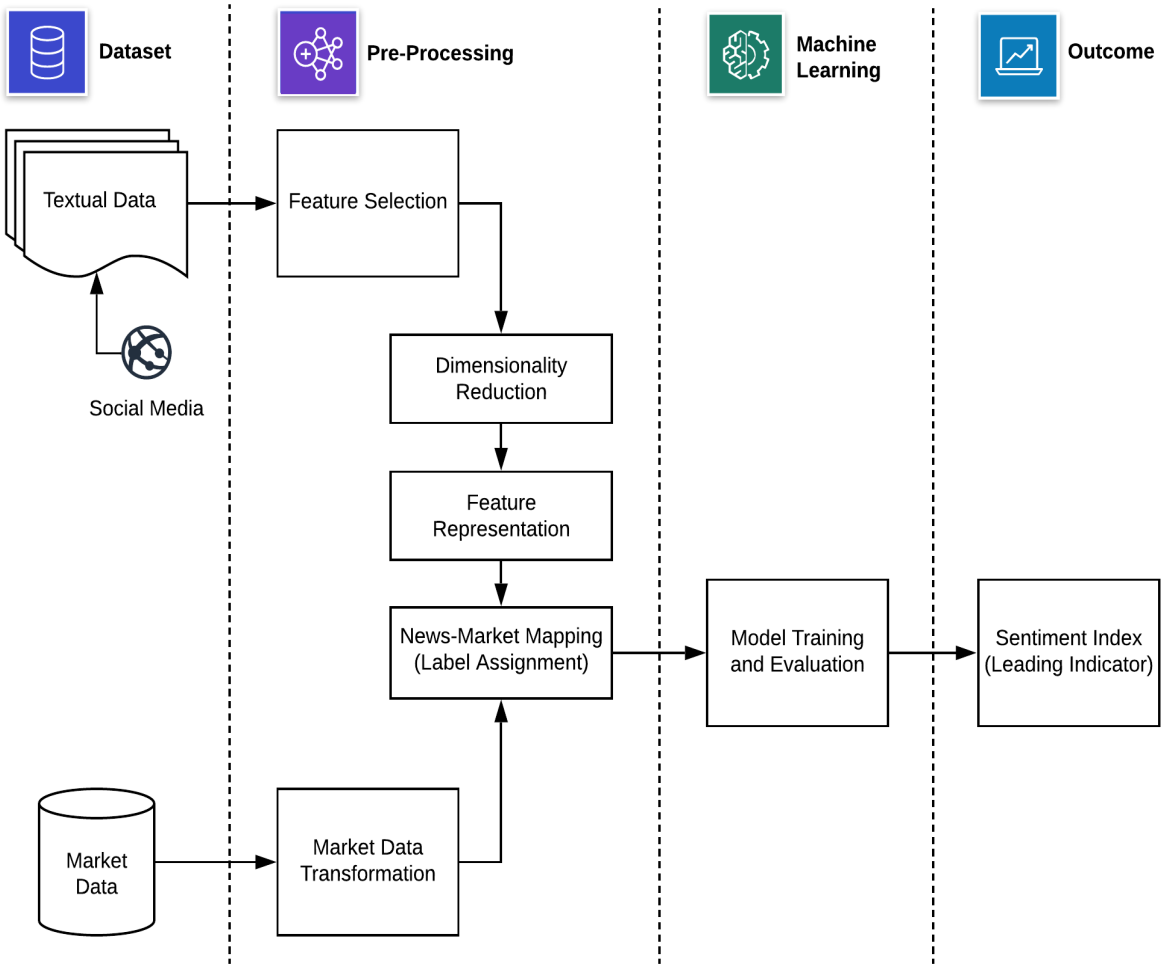
The daily data on SET50 index is chosen for this project for a number of reasons. First, the SET50 index has co-moved with the SET index and the SET100 index at least since 2007. Second, the search for tonal words in social media and collecting data on these words for our analysis are time-consuming even for an individual company. With time and budget limitation, the number of listed companies, to which Thai social media refer, in this project is limited to 50 largest and most liquid listed companies in the SET.

Unlike existing studies which use English words from particular dictionaries (see e.g. Fraiberger, Lee, Puy and Ranciere (2018)), we investigate tonal words in Thai; these Thai words are frequently seen in Thai media posts. Research in market sentiment for emerging markets focuses on announcements by firms, tonal words in economic news or policy releases or some investment sentiment indexes which tend to be computed from macroeconomic indicators. Differently, in this project, we analyze the effect of tonal words in Thai online and social media which show public emotion and perception toward companies listed in the SET50 index. Thai social media platforms are selected for this project, based on their popularity (e.g. visibility in online community, or number of LIKEs or subscribers) and relevance (e.g. specializing in providing financial and investment decision).

These tonal words can reflect investor sentiment toward the firms. Machine Learning algorithms are used to conduct a bag-of-word analysis which will provide some statistics of tonal

words for 50 firms over the period 2015-2018 and these statistics are used to create a new leading indicator which reflects investors' sentiment toward the 50 firms listed in SET50. Figure 3 summarized a process flow for constructing our new leading indicator which hereafter will be called a sentiment index. The process consists of 2 main stages: 1) machine learning analysis and 2) sentiment index construction. The process of the sentiment index construction is detailed in the following sub-section.

Figure 3: Process flow of the sentiment analysis for a sentiment index



3.3 Sentiment Index Construction

Sentiment analysis, also known as opinion mining, is a study in natural language processing (NLP) to analyze investors' opinions, attitudes, and emotions towards the stock market. As shown

in Figure 3, this project extract tonal words from social media and these will be analyzed and used to construct the sentiment index. The first stage known as “machine learning analysis” includes *datasets preparation*, *pre-processing*, and *NLP* (machine learning). The second stage employs the statistics obtained from the first stage to construct a sentiment index.

Machine Learning Analysis

1) Dataset Preparation

Datasets preparation is a process of textual data collection using Scrapy, aiming at collecting daily news related to all SET50 companies over the period between 2015—2018. Note that the firms of which online news and posts are retrieved from social media change every six months because the list of SET50 constituents changes semi-annually. The workflow for the textual data collection is shown in Figure 1.

A word bank, which is composed of an initial set of keywords (stock-related words and tonal words) suggested from experts, are created. Stock-related keywords are used in searching and filtering SET50 news from the selected 10 online social media. This is important due to the tons of the SET50 related information that can be found during the search, which potentially causes a high-power consumption and complexity in calculating the sentiment index.

In particular, the word bank is used in the web spider to retrieve online posts and news from the selected Thai online social media listed in Section 3.1 and the results (online posts and news) are returned in a form of Hypertext Markup Language (HTML). Later, some important information are extracted into a structured form e.g. *date*, *title*, *body*, *source*, and *tag* (if available), and then stored into our sentiment database.

In this stage, a dictionary of tonal words (e.g. positive, negative, and neutral words) and stock-related words is developed and updated. These tonal words are analyzed; directional words are identified and used to indicate the tone of context words. Table 1 shows examples of initial tonal words from our sentiment dictionary.

Table 1: Example of included Thai tonal words with their pronunciation and English translation

Positive words	Neutral words	Negative words
กำไรโต kam-ri-to growth profits	ทรงตัว shong-tua settled	กำไรหด kam-ri-hod shrink profits
กระชากขึ้น kra-chak-khuan snatched up	ลุ้นต่อ lun-tor keep looking forward	ถอยแรง thoy-rang drastically down
ชนะคดี cha-na-kha-di win the case	อย่าเพิ่งให้น้ำหนักมาก yaah-pueang-hi-num-nak-mak don't give too much weight)	เทขาย tay-khay sell a stock
ทำนิวไฮ tum-new-high reached a new high	นิ่งไว้ ning-wi don't do anything	ทำนิวโลว์ tum-new-low caused a new low
โดดเด่น dod-den outstanding	ประคองตัว pra-khong-tua stable	กำไรลดฮวบ kam-ri-lod-hwuab drastically reduced profits)

2) Pre-processing

Computing textual information can be difficult compared with mathematics that can be done with operators. Working on textual information requires a number of steps, such as *word segmentation*, *tokenization*, *stop words filtering*, *negation handling*, and *normalization*. These procedures are implemented to create text data representation for machine which, in this project, a bag-of-word analysis was used (see Figure 4). These pre-processing procedures are important; they allows machine to understand the textual information in natural language computation.

The bag-of-word (BoW) analysis is a method to extract features from textual information. These features can be used for training machine learning algorithms. It creates a set of unique

vocabulary occurring in all documents. Each sentence is segregated into words and the frequency of each word is calculated regardless of the word's order of appearance in the sentence.

Before the creation of BoW, a number of steps were performed in following order.

- ***Word Segmentations***: a process that identifies the end of each sentence as, unlike Western languages, Thai language does not have any punctuation between clauses and a "full stop" mark. This can cause many problem in a BoW analysis. For example, some words are intended to be in the first sentence, but without a punctuation or mark they can be mistaken to be included in the second sentence. In an extreme case, some words can be interpreted differently if the sentence is incorrectly segmented.

  For example, "แม่น้ำน่าน" (*mae-nam-nan*, Nan river), it is a name of the river. If it is segmented into 3 words, it will be interpreted differently.

  - แม่ (*mae*) -- mother
  - น้ำ (*nam*) -- water
  - น่าน (*nan*) – Nan (name entity)

- ***Tokenization***: a process that breaks up a sequence of strings into small pieces such as words, keywords, phrases, and other elements. Tokens can be individual words, phrase or even whole sentences. In the tokenization process, punctuation and full stops marks are not included.

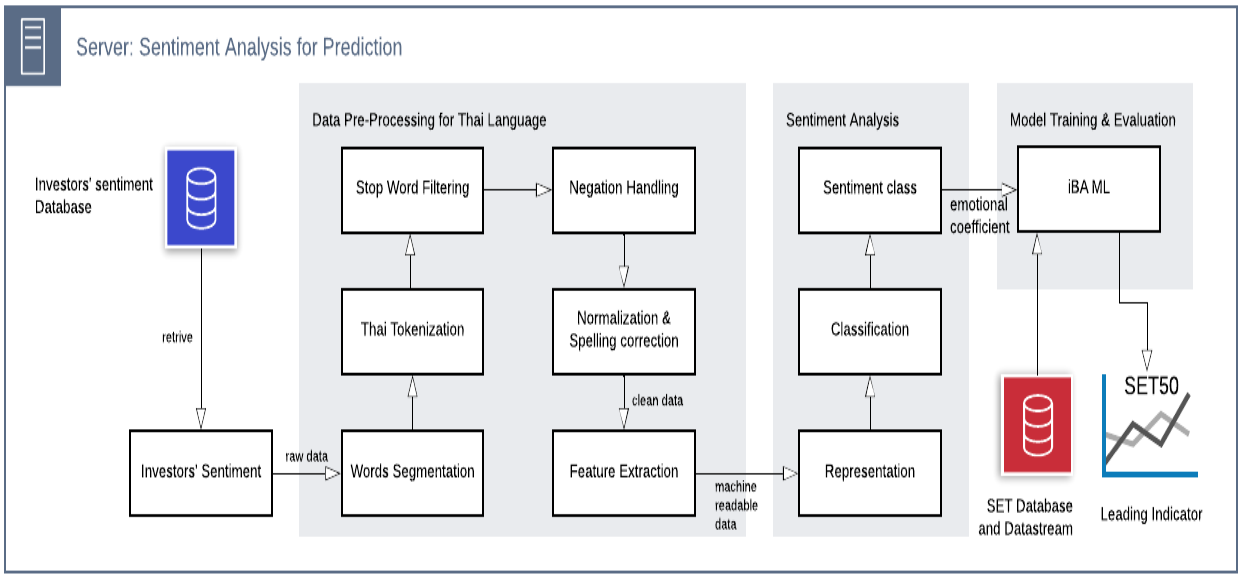
- ***Stop words filtering***: a process that filters out the words that have no meaning or do not contain enough significance to be used in the sentiment analysis.

- ***Negation handling*****:** a process that groups up the negative words into one word. For example, "ไม่ดี" (*mi-dee*, no good/ not good/ not okay) can be split into 2 words of "ไม่" (*mi*, no/ not/ reject/ refuse) and "ดี" (*dee,* good, nice, brilliant, cool). If the machine does not recognize this combination of words, then the machine may read this word as two separate words and incorrectly analyze it as one positive word and one negative word.

- ***Normalization*****:** a process that convert words to root words, for example "good" is a root word for excellent, fine; valuable; desirable, favorable, beneficial, etc. With this, it can increase an accuracy in the sentiment classification.

- ***Feature extraction***: a process that segregate a sentence into a collection of unique words and counts the frequency of each word regardless of the order in which they appear in the sentence. At the end of this process, the BoW (representation) is created.

Figure 4: Sentiment Analysis and Prediction Workflow



3) NLP

Towards the sentiment classification, in this project, sets of positive, negative, and neutral words were created and denoted as W'_{pos}, W'_{neg} and W'_{neu}, respectively. The appearance of each word in a post is counted, as can be denoted by W = {w_i,n_i}. For example, a set of positive words, W'_{pos}, contains words such as "โดดเด่น"/outstanding (3 appearance) and "กำไรโต"/growing profits (2 appearance), i.e. W'_{pos} = {(โดดเด่น,3), (กำไรโต,2)}.

To determine a mood or sentiment (S) score of the given document, parameters of and were fed in the sentiment analysis process using a sentiment analysis equation given by Ren et al. (2018), as follows:

$$S = sentiment(W'_{pos}, W'_{neg}, W'_{neu}) = \begin{cases} \dfrac{2 \times W'_{pos}}{(W'_{pos} + W'_{neg}) - 1} & , \quad W'_{pos} > W'_{neg} \\ 0 & , \quad W'_{pos} = W'_{neg} \\ \dfrac{1 - 2 \times W'_{neg}}{(W'_{pos} + W'_{neg})} & , \quad W'_{pos} < W'_{neg} \end{cases}$$

- Eq(1)

In the end, the sentiment score was later classified using following conditions in Table 2.

Table 2: Sentiment Classification

Sentiment class	Condition
Positive	S > 0.05
Neutral	-0.05 ≤ S ≤ 0.05
Negative	S < -0.05

Sentiment index construction

To construct the sentiment index, we employ the percentage of positive tonal words and the percentage of negative tonal words for each of the 50 listed firms obtained from the bag-of-word analysis in the previous stage to calculate the weighted average value of positive sentiment (positive market sentiment index) and the weighted average value of negative sentiment (negative market sentiment index). The weight for each firm is time-varying; it is calculated from dividing the market value of each firm with the total market value of all 50 firms listed on the SET50 index. The calculation of weight is done for each calendar day because media articles can be published even on non-trading days. The data on market values is available only on trading day; therefore, we assume that the market values on the day after weekend or holiday are the same as the values on the day before weekend or holiday.

Next, we calculate a composite market sentiment index which is defined as the difference between the positive market sentiment index and negative market sentiment index.

3.4 Relationship between SET50 and Sentiment Index

After employing a bag-of-word analysis to extract tonal words and analyze the sentiment in online posts and constructing a composite market sentiment index, a relationship between the SET50 index returns and the market sentiment index is examined using a multiple regression. The study on sentiment effects in stock returns tends to employ daily data. One reason is that the stock market operates daily, allowing investors to react to news immediately or on the next few days

and thus stock prices to absorb information quickly. Another is that employing weekly or less frequent data will neglect any existing short-life effect of market sentiment.

A significant relationship between the composite market sentiment index and the movement of the SET50 index will imply that this newly introduced sentiment index can act as a new leading indicator for the SET.

The general form of regression is

$$Y = C + AX + BZ + U \qquad \text{— Eq (2)}$$

The vector of dependent variables (Y) are SET50 index returns (SET50(t)), cumulative SET50 index returns between day $t-1$ and *t* (SET50(t-1,t)), and cumulative SET50 index returns between day $t-1$ and $t+1$ (SET50(t-1,t+1)). The vector of independent variables in our interest (X) includes the current, past and cumulative values of the composite sentiment index. The sentiment(t) variable denotes the value of the composite market sentiment index at time t. Note that the observations of cumulative index returns only include observations on trading days while the calculation of cumulative value for the composite market sentiment index includes the values on non-trading day if t-1 or t+1 is non-trading day ; this allows us to include articles published during non-trading days in our investigation.

The regression includes a vector of some control variables (Z). These control variables are local market variables (e.g. log-trading value which proxies for market liquidity, foreign investors' net trade value which reflects the amount of foreign capital flows into the market and local individual investors' net trade value reflecting domestic individual investment) and global market variables (e.g. S&P global index returns and VIX). We also include a set of dummy variables which proxy for day and month anomaly effects. The "Monday" dummy variable has a value of 1 for the trading day that is on Monday and 0 for otherwise while the "Friday" dummy variable has a value of 1 for the trading day that is on Friday and 0 for otherwise. As Thailand has many non-weekend holiday, we also introduce a "Holiday" dummy variable which has a value of 1 for the first day after holiday or weekend and 0 for otherwise to control the holiday effect. Another dummy variable is "January" which has a value of 1 for the observations in January.

Using daily data limits the choice of macroeconomic variables to be included in the regression. We first attempt to include the interest rate; however, the exchange rate system of Thailand is the managed float exchange rate regime which is consistent with inflation-targeting regime so the currency value already reflects the overall economic performance and closely links with the

interest rate. Including the interest rate and the exchange rate of Thai Baht in the same regression can cause weakly multi-collinearity problem and underestimate their effects, so here we employ the daily exchange rate only. Given that Thailand is an exporting country, the value of Thai Baht (THB) would be an important macroeconomic variable.

Chapter 4 Finding and Discussion

4.1 Sentiment Index and Descriptive Statistics

The descriptive statistics of SET50 index, sentiment index and control variables are reported in Table 3. In order to avoid the spuriousness of regression using non-stationary variables, our variables in levels are tested for stationarity and are transformed to remove the unit root or trend. More specifically, the daily SET50 index is transformed to be SET50 index returns (in percentage). With regard to market variables which reflect activities in the overall SET market, the total market trade value is in logarithmic form and the ratio of purchase to sell is used for both foreign investors and local individual investors instead of their net value.

We note that local institutional trade in the SET is not included as a control variable because adding it to the model causes a multi-collinearity problem and, when including it, the coefficient for local institutional trade is insignificant. In addition, the total values of trading by these two investor types are much higher than the total values of trading by local institutional investors and proprietary traders. Employing the procedure explained in Section 3.3, we calculate the positive market sentiment index which measures how investors are optimistic toward the market and the negative market sentiment index which measures how investors are pessimistic toward the market. Figure 5 presents daily positive and negative sentiment indexes. Figure 5 highlights that during 2015-2018 the market sentiment tend to be more positive than negative, especially in 2017-2018 when the THB currency is appreciated.

Then, we calculate the composite market sentiment index which concludes whether the market sentiment is on average positive or negative. The daily market sentiment index over the period 2015-2018 is plotted in Figure 6. The higher the market sentiment index, the more optimistic the market.

Table 3: Descriptive Statistics

Variables	Mean	Median	Max	Min	Std.	Skew.	Kurt.	Obs.
SET50(t)	0.008	0.020	4.090	-5.080	0.878	-0.088	5.735	976
SET50(t-1,t)	0.017	0.020	5.260	-5.500	1.242	0.019	4.594	975
SET50(t-1,t+1)	0.027	-0.030	5.560	-6.500	1.525	-0.032	4.154	974
Sentiment(t)	0.147	0.137	0.858	-0.891	0.228	-0.209	3.671	1458
Ln(Total Trade Value(t))	0.147	0.137	0.858	-0.891	0.228	-0.208	3.673	976
Ratios of Buy(t)/Sell(t):	0.293	0.294	1.384	-1.060	0.355	-0.070	2.975	
Local Institution	1.154	1.084	8.150	0.248	0.493	3.974	46.196	976
Local Individual	1.006	0.999	1.581	0.532	0.104	0.453	5.736	976
Foreign Institution	0.981	0.969	1.793	0.586	0.138	1.147	7.543	976
Proprietary Traders	1.018	1.007	1.766	0.248	0.172	0.574	4.779	976
Other control variables:	24.071	24.055	25.216	22.939	0.341	0.154	3.197	
S&P Global Return(t)	0.068	0.063	7.858	-6.386	1.383	-0.145	7.676	975
VIX	15.119	13.755	40.740	9.140	4.689	1.507	5.857	976
%Change in THB/USD	-0.002	-0.011	1.168	-1.362	0.276	0.010	4.609	975
Trade values (mil. THB):								
Local Institution Buy(t)	5450	5070	21100	1370	2210	1.616	8.369	976
Local Institution Sell(t)	5070	4680	22000	1210	2120	1.885	10.802	976
Local Institution Total(t)	10527	10028	29327	2974	3738	1.276	2.950	976
Local Institution Net(t)	378	367	15700	-16500	2190	0.102	12.327	976
Local Individual Buy(t)	25600	24100	71200	10800	7740	1.137	5.226	976
Local Individual Sell(t)	25600	24200	66700	10900	7810	1.061	4.664	976
Local Individual Total(t)	51117	48428	137935	21711	15284	1.074	1.828	976
Local Individual Net(t)	1	-17	17700	-13600	2850	0.122	8.318	976
Foreign Institution Buy(t)	14700	13300	52100	1210	6130	1.203	5.886	976
Foreign Institution Sell(t)	15100	13600	59200	1530	6330	1.315	6.394	976
Foreign Institution Total(t)	29726	26935	110594	2731	12299	1.260	3.139	976
Foreign Institution Net(t)	-400	-394	10700	-10600	2000	0.068	6.255	976
Proprietary Traders Buy(t)	5490	5130	17300	1160	1950	1.268	6.040	976
Proprietary Traders Sell(t)	5470	5070	18300	1610	1980	1.553	7.401	976
Proprietary Traders Net(t)	21	32	3370	-4810	918	-0.519	5.860	976

Figure 5: Daily positive and negative sentiment indexes

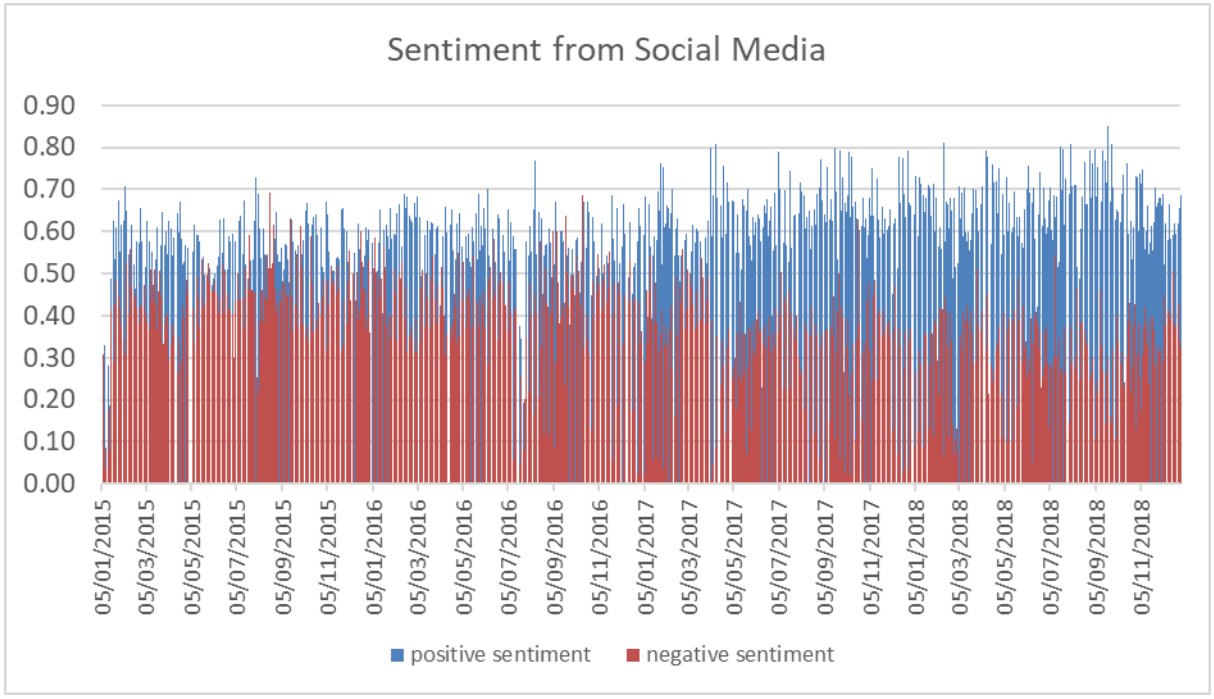
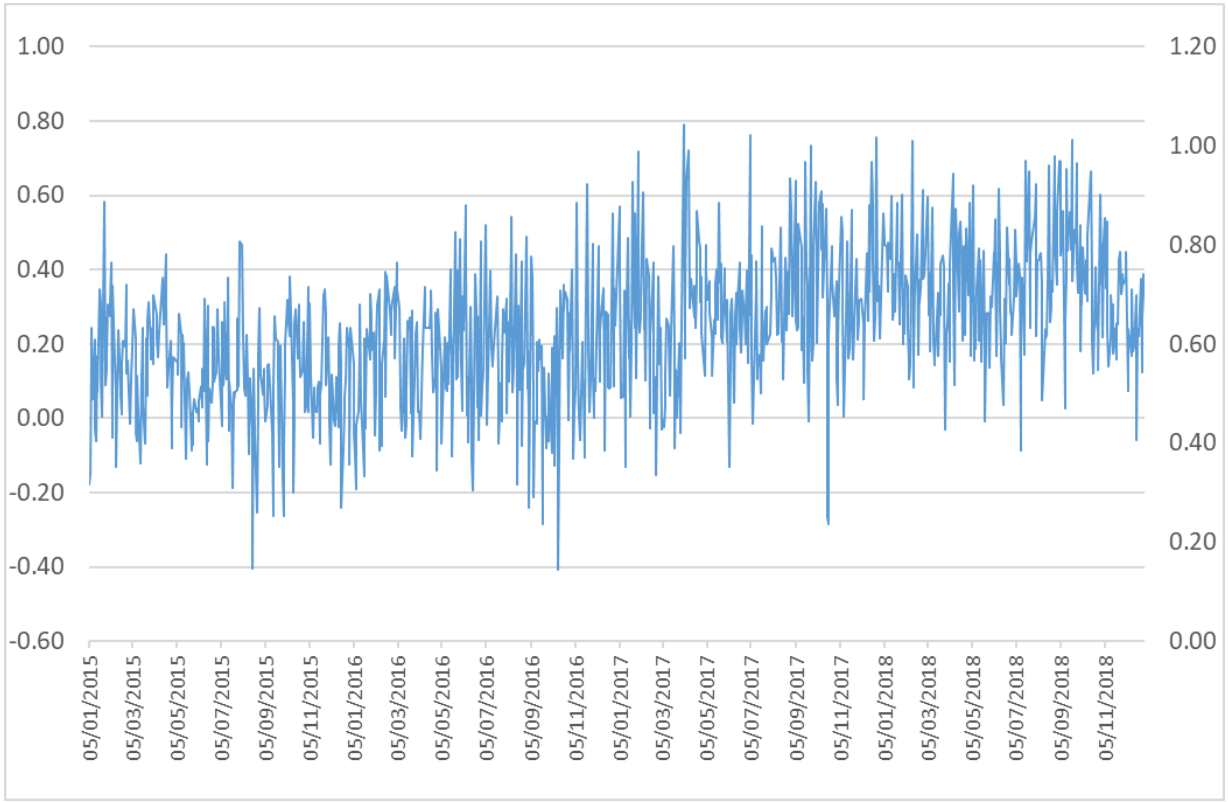


Figure 6: Daily market sentiment index



4.2 Correlation Analysis

We first test the correlation between SET50 index returns and the sentiment index. Both current and cumulative values of the SET50 index return and sentiment index are investigated. An inspection of Table 4 highlights that the current value of sentiment index, “Sentiment(t)” is significantly and positively correlated with the current SET50 index return (SET50(t)) and the cumulative 2-day SET50 index returns between t-1 and t (SET50(t-1,t)) and the cumulative 3-day SET50 index returns between t-1 and t+1 (SET50(t-1,t+1)). In addition, the sentiment on the previous day “Sentiment(t-1)” and the cumulative past sentiment “Sentiment(t-2,t-1)” have a significant positive relationship with SET50(t-1,t) and SET50(t-1,t+1).

Table 4: Correlation Analysis

	SET50(t)	SET50(t-1,t)	SET50(t-1, t+1)	Sentiment(t)	Sentiment(t-1)
SET50(t-1,t)	**0.7079**				
	(31.2489)				
SET50(t-1, t+1)	**0.5774**	**0.8180**			
	(22.0476)	(44.3423)			
Sentiment(t)	**0.3664**	**0.3845**	**0.3198**		
	(12.2789)	(12.9850)	(10.5230)		
Sentiment(t-1)	0.0229	**0.2111**	**0.1715**	**0.2177**	
	(0.7134)	(6.7331)	(5.4280)	(6.9529)	
Sentiment(t-2,t-1)	-0.0107	**0.1072**	**0.0899**	**0.1498**	**0.7695**
	(-0.3332)	(3.3628)	(2.8128)	(4.7244)	(37.5657)

Note: Correlation coefficients are emboldened if the probability of t-statistics for testing whether the correlation between the pair of variables is significant at 0.05 significance level. The values in parentheses are t-statistics.

4.3 Regression Analysis

SET50 index return day t

Next, we test whether our sentiment index has any relationship with SET50 index return if controlling for other effects. Table 5 reports the regressions of SET50 index return on our new sentiment index and some control variables. Eight different models are estimated to investigate the power of different sets of variables.

Models (1), (4), (7) and (8) in Table 5 confirm that there are insignificant calendar effects on SET50 index return. The models in Table 5 consistently report that an increase in total trade value in the stock market is related to a higher SET50 index return. Models (2)-(4), (7) and (8) suggest that trade by local individual investors and foreign investors are significantly related to the SET50 index return while trade by foreign investors has less impact than local individual investors. The negative sign of the coefficients for trade by local individual investors and foreign investors indicates that the gain in SET50 index is related to an increase in purchase by other traders than these two types of investors (local institutional investors and proprietary traders) because more selling or less buying by foreign and local individual investors, while the total trade value in the market is unchanged, implies more buying or less selling by other investors. Moreover, the models shows that an increase in the SET50 index return is associated with an increase in the S&P global index return and the appreciation of Thai Baht on the same day.

Model (8) is selected as the best fitted model because it has the greatest explanatory power and the lowest value for all three information criteria. The likelihood ratio for Model (8) testing whether the sentiment index variables are jointly redundant are equal to 69.49 (prob. = 0.00), indicating that the sentiment index variables are significantly associated with the SET50 index return and should be included in the model. According to the estimate of Model (8), if investor sentiment toward the firms was not good in the past few days but becomes more positive today, then the SET50 index will rise.

Cumulative SET50 index return between day t-1 and day t

Turning to the cumulative index return, we investigate whether this sentiment index has any relationship with cumulative SET50 index return. Table 6 reports the regressions of cumulative 2-day SET50 index return (t-1,t) on sentiment index variables (t, t-1 and between t-3 and t-2) when the value of control variables included are the values on day t. Additional eight different models are estimated to investigate the power of different sets of variables. The estimates of models (1'), (4'), (7') and (8') in Table 6 indicate that there are insignificant calendar effects on the cumulative SET50 index return. Like the results for the current SET50 index return, the cumulative SET50 index return is positively associated with the total trade value in the stock market.

Consistent with the results shown in Table 5, all models in Table 6 suggest that local individual investors' trade plays a more important role than foreign investors' trade. The negative sign of the coefficients for trade by local individual investors and foreign investors indicates that the cumulative gain in SET50 index is related to an increase in purchase by other traders than foreign investors and local individual investors because, given an unchanged trade value in the market, more selling or less buying by foreign investors and local individual investors implies more buying or less selling by the other investors.

Model (7') and Model (8') are obviously the two best fitted models because they have significantly higher explanatory power (R-square value = 77%) and lower values for all three information criteria than other models. Both Model (7') Model (8') show that an increase in cumulative SET50 index return (t-1,t) is associated with an increase in the S&P global index return and the appreciation of Thai Baht on day t. Given that the sentiment(t-1) and sentiment(t-3,t-2) variables in Model (8') have significant coefficients, the estimation results suggest that all three sentiment index variables are significantly associated with the SET50 index return. According to the estimate of Model (8'), if investor sentiment toward the firms was not good in the past three days but becomes more positive today, then the cumulative SET50 index will rise.

Cumulative SET50 index return between day t-1 and day t+1

Table 7 reports the regressions of cumulative 3-day SET50 index return (t-1,t+1) on sentiment index variables (t, t-1 and between t-3 and t-2) when the value of control variables included are the values on t. The estimates of models (1''), (4''), (7'') and (8'') in Table 7 indicate that there are insignificant calendar effects on the cumulative SET50 index return.

Models (2'')-(4'') show significant relationships between cumulative SET50 index returns and trade by local individual investors as well as total trade in the market and Models (5'') and (6'') report significant sentiment effects, but Models (7'') and (8''), which have higher R-squared values (68%), report insignificant trade and sentiment effect. In addition, we find that the cumulative SET50 index return is positively associated with the S&P global index return on day t, but negatively associated with the percentage change in the value of Thai Baht on day t. In order to correct the remaining autocorrelation problem, the cumulative 3-day SET50 index return (t-2,t) and the SET50 index return on day t-2 are added; the result shows that the former has a positive impact on the cumulative 3-day SET50 index return (t-1,t+1) while the latter has a negative impact.

TGARCH (1,1,1) models

As shown in Tables (5)-(7), all models have clustering volatility of residuals, the first order of Threshold Generalised Autoregressive Conditional Heteroscedasticity (T-GARCH (1,1,1)) is estimated for all three dependent variables where the mean equations of TGARCH models are Model (8) in Table 5, Model (8') in Table 6 and Model (8'') in Table 7. The estimates of TGARCH models are reported in Table 8. Although the leverage effect on SET50 index return volatility is detected, the explanatory power is not improved.

The estimates of mean equation is unchanged after correcting the clustering volatility problem. The significance and sign of coefficients in Table 8 are the same as those in Model (8) in Table 5, Model (8') in Table 6 and Model (8'') in Table 7. Therefore, our findings confirm that this new sentiment index has significant relationships with the SET50 index returns and the sentiment index we introduce can suggest the change in SET50 index. The sentiment index

can signal the SET50 index up to the next 3 days. Because the co-movement between SET50 index and SET index, this sentiment index can therefore also signal the movement of SET index.

Table 5: The regression of SET50 index return on day t

	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
Variable	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Constant	-0.002	0.035	1.333	1.628	1.287	1.798	1.437	1.910	**-0.350**	0.044	**-0.168**	0.056	3.039	1.800	2.264	1.785
Monday(t)	0.036	0.191					-0.017	0.118					-0.049	0.115	-0.091	0.113
Friday(t)	0.087	0.073					0.052	0.053					0.040	0.051	0.029	0.050
Holiday(t)	-0.123	0.179					0.019	0.111					-0.077	0.110	-0.167	0.113
January(t)	0.169	0.123					0.106	0.086					0.115	0.086	0.109	0.084
Foreign(t)			**-1.059**	0.233	**-0.939**	0.247	**-0.943**	0.245					**-0.792**	0.244	**-0.764**	0.241
Local individual(t)			**-6.000**	0.508	**-6.410**	0.551	**-6.400**	0.551					**-5.869**	0.552	**-5.733**	0.550
Ln(Trade value(t))			**0.237**	0.057	**0.252**	0.063	**0.245**	0.068					**0.139**	0.068	**0.170**	0.068
S&P global return(t)			**0.083**	0.020	**0.078**	0.020	**0.078**	0.020					**0.075**	0.019	**0.073**	0.019
VIX(t)			0.002	0.005	0.001	0.006	0.001	0.006					0.007	0.006	0.003	0.006
FX return(t)			**-0.252**	0.096	**-0.257**	0.101	**-0.249**	0.102					**-0.198**	0.095	**-0.197**	0.093
SET50(t-1)					**-0.198**	0.029	**-0.199**	0.029	**-0.118**	0.040	-0.058	0.043	**-0.212**	0.028	**-0.192**	0.031
Sentiment(t)									**1.903**	0.190	**2.453**	0.220	**0.882**	0.144	**1.018**	0.155
Sentiment(t-1)											**-0.462**	0.218			-0.130	0.153
Sentiment(t-2,t-1)											**-0.441**	0.135			**-0.185**	0.090
	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.
R-squared	0.007		0.475		0.510		0.512		0.145		0.194		0.535		0.546	
Adjusted R^2	0.003		0.472		0.507		0.506		0.143		0.190		0.529		0.539	
S.E. of regression	0.876		0.638		0.616		0.616		0.793		0.770		0.602		0.596	
Log likelihood	-1253.6		-941.1		-907.1		-905.5		-892.0		-869.6		-881.7		-870.7	
F-stats	1.773	0.132	145.967	0.000	143.959	0.000	91.827	0.000	63.491	0.000	45.057	0.000	92.309	0.000	82.305	0.000
Wald F-stats	1.517	0.195	47.648	0.000	42.743	0.000	29.520	0.000	50.296	0.000	31.975	0.000	39.466	0.000	35.918	0.000
AIC	2.579		1.945		1.877		1.882		2.377		2.323		1.835		1.817	
SIC	2.604		1.980		1.917		1.942		2.396		2.354		1.900		1.892	
HQC	2.589		1.958		1.892		1.905		2.384		2.335		1.860		1.845	
DW stat	2.099		2.308		1.931		1.928		1.823		2.022		1.930		1.966	
Q(1)	2.055	0.152	23.272	0.000	1.146	0.284	1.248	0.264	3.449	0.063	0.0777	0.781	1.183	0.277	0.280	0.597
Q(2)	2.288	0.319	25.850	0.000	1.808	0.405	1.770	0.413	4.646	0.098	1.2703	0.530	2.347	0.309	1.022	0.600
Q(8)	8.726	0.366	29.949	0.000	4.776	0.781	5.955	0.744	9.894	0.273	2.2763	0.971	6.113	0.635	7.771	0.456
$Q^2(1)$	14.583	0.000	14.044	0.000	3.1101	0.078	3.147	0.076	12.706	0.000	14.036	0.000	5.270	0.022	4.994	0.025
$Q^2(2)$	23.697	0.000	22.410	0.000	9.9255	0.007	10.550	0.005	17.864	0.000	20.094	0.000	13.472	0.001	11.200	0.004
$Q^2(8)$	83.037	0.000	50.362	0.000	28.699	0.000	29.862	0.000	28.105	0.000	28.039	0.000	40.110	0.000	34.344	0.000

Note: The emboldened coefficients are significantly different from zero at a 0.05 significance level. The probability of Q-statistics for residuals that is greater than 0.05 indicates the existence of autocorrelation in residuals while the probability of Q-statistics for squared residuals that is greater than 0.05 indicates the clustering volatility.

Table 6: The regression of accumulated SET50 index return between day t-1 and day t

	(1')		(2')		(3')		(4')		(5')		(6')		(7')		(8')	
Variable	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Constant	-0.017	0.063	1.102	2.928	1.447	2.174	1.372	2.301	**-0.453**	0.048	**-0.349**	0.060	3.039	1.820	1.933	1.806
Monday(t)	0.182	0.274					0.030	0.146					-0.047	0.116	-0.027	0.114
Friday(t)	0.040	0.099					-0.002	0.073					0.042	0.051	0.049	0.049
Holiday(t)	-0.172	0.259					0.037	0.136					-0.081	0.111	-0.160	0.114
January(t)	0.3738	0.2039					0.114	0.104					0.122	0.089	0.112	0.087
Foreign(t)			-0.452	0.369	**-0.996**	0.314	**-0.999**	0.315					**-0.760**	0.246	**-0.747**	0.241
Local individual(t)			**-8.072**	0.694	**-7.379**	0.626	**-7.385**	0.628					**-5.860**	0.554	**-5.736**	0.552
Ln(Trade value(t))			**0.313**	0.107	**0.287**	0.078	**0.290**	0.083					**0.138**	0.069	**0.185**	0.068
S&P global return(t)			**0.057**	0.026	**0.077**	0.025	**0.078**	0.025					**0.074**	0.019	**0.074**	0.019
VIX(t)			-0.004	0.009	0.003	0.007	0.002	0.007					0.006	0.006	0.002	0.006
FX return(t)			-0.278	0.160	-0.194	0.129	-0.182	0.131					**-0.205**	0.095	**-0.203**	0.094
SET50(t-2,t-1)					**0.369**	0.024	**0.367**	0.024	**0.457**	0.024	**0.467**	0.024	**0.786**	0.029	**0.803**	0.031
SET50(t-2)													**-0.822**	0.042	**-0.819**	0.043
Sentiment(t)									**2.030**	0.183	**2.298**	0.199	**0.886**	0.146	**1.052**	0.162
Sentiment(t-1)											-0.010	0.170			**-0.317**	0.122
Sentiment(t-3,t-2)											**-0.515**	0.104			**-0.158**	0.055
	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.
R-squared	0.008		0.467		0.589		0.590		0.351		0.372		0.770		0.774	
Adjusted R^2	0.004		0.463		0.586		0.586		0.349		0.370		0.767		0.770	
S.E. of regression	1.240		0.910		0.798		0.799		1.001		0.985		0.599		0.595	
Log likelihood	-1590.6		-1287.9		-1158.7		-1157.5		-1381.7		-1365.2		-875.5		-867.9	
F-stats	1.923	0.104	141.231	0.000	197.874	0.000	125.955	0.000	262.136	0.000	143.658	0.000	247.707	0.000	218.577	0.000
Wald F-stats	0.985	0.415	59.154	0.000	104.349	0.000	72.206	0.000	355.545	0.000	188.876	0.000	170.853	0.000	165.575	0.000
AIC	3.273		2.656		2.396		2.401		2.843		2.813		1.826		1.815	
SIC	3.298		2.691		2.436		2.462		2.858		2.839		1.897		1.895	
HQC	3.283		2.670		2.411		2.424		2.849		2.823		1.853		1.846	
DW stat	0.998		1.485		1.849		1.844		1.588		1.649		1.917		1.970	
Q(1)	243.390	0.000	64.556	0.000	5.322	0.021	5.710	0.017	40.504	0.000	29.368	0.000	1.555	0.212	0.174	0.677
Q(2)	244.140	0.000	66.221	0.000	36.846	0.000	37.684	0.000	111.230	0.000	115.370	0.000	5.147	0.076	2.327	0.312
Q(8)	254.130	0.000	72.229	0.000	38.114	0.000	38.955	0.000	119.620	0.000	118.130	0.000	9.371	0.312	8.959	0.346
$Q^2(1)$	84.316	0.000	41.391	0.000	7.7524	0.005	7.697	0.006	5.661	0.017	4.8813	0.027	4.259	0.039	5.181	0.023
$Q^2(2)$	96.300	0.000	60.471	0.000	57.532	0.000	57.759	0.000	50.569	0.000	40.918	0.000	11.337	0.003	12.397	0.002
$Q^2(8)$	205.220	0.000	132.070	0.000	95.05	0.000	95.478	0.000	134.090	0.000	115.93	0.000	34.746	0.000	35.101	0.000

Note: The emboldened coefficients are significantly different from zero at a 0.05 significance level. The probability of Q-statistics for residuals that is greater than 0.05 indicates the existence of autocorrelation in residuals while the probability of Q-statistics for squared residuals that is greater than 0.05 indicates the clustering volatility.

Table 7: The regressions of accumulated SET50 index return between day t-1 and t+1

	(1'')		(2'')		(3'')		(4'')		(5'')		(6'')		(7'')		(8'')	
Variable	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Constant	0.012	0.080	-1.765	3.685	-1.980	2.642	-1.592	2.750	**-0.201**	0.060	-0.093	0.072	-2.194	2.125	-2.488	2.150
Monday(t)	0.069	0.303					-0.103	0.188					-0.129	0.148	-0.120	0.148
Friday(t)	-0.057	0.109					-0.108	0.089					-0.073	0.068	-0.066	0.072
Holiday(t)	-0.144	0.278					0.125	0.155					0.073	0.119	0.064	0.132
January(t)	0.557	0.272					0.170	0.135					0.150	0.094	0.148	0.095
Foreign(t)			-0.196	0.458	-0.374	0.331	-0.381	0.336					0.310	0.236	0.311	0.236
Local individual(t)			**-7.907**	0.755	**-4.215**	0.597	**-4.245**	0.604					-0.171	0.406	-0.156	0.402
Ln(Trade value(t))			**0.412**	0.142	**0.269**	0.111	**0.255**	0.115					0.084	0.095	0.097	0.095
S&P global return(t)			**0.142**	0.038	**0.129**	0.033	**0.130**	0.033					**0.086**	0.025	**0.086**	0.025
VIX(t)			0.001	0.013	0.008	0.008	0.007	0.008					0.003	0.007	0.002	0.007
FX return(t)			**-0.495**	0.169	**-0.259**	0.127	-0.245	0.127					**-0.221**	0.111	**-0.221**	0.112
SET50(t-2,t-1)					**0.471**	0.032	**0.469**	0.033	**0.611**	0.026	**0.614**	0.027	**0.955**	0.033	**0.954**	0.033
SET50(t-2)													**-1.025**	0.053	**-1.020**	0.053
Sentiment(t)									**0.927**	0.210	**1.197**	0.225	0.046	0.177	0.079	0.189
Sentiment(t-1)											-0.109	0.192			-0.025	0.175
Sentiment(t-3,t-2)											**-0.468**	0.114			-0.058	0.085
	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.	Stats	Prob.
R-squared	0.011		0.352		0.494		0.496		0.435		0.447		0.683		0.683	
Adjusted R^2	0.007		0.348		0.491		0.491		0.433		0.444		0.679		0.678	
S.E. of regression	1.520		1.232		1.089		1.089		1.149		1.137		0.865		0.866	
Log likelihood	-1787.5		-1581.7		-1459.6		-1457.7		-1513.9		-1503.4		-1232.7		-1232.4	
F-stats	2.657	0.032	87.445	0.000	134.786	0.000	86.119	0.000	372.888	0.000	195.436	0.000	158.844	0.000	137.482	0.000
Wald F-stats	1.195	0.311	49.370	0.000	127.742	0.000	81.331	0.000	393.756	0.000	220.307	0.000	181.2722	0.000	158.874	0.000
AIC	3.681		3.262		3.017		3.021		3.118		3.101		2.563		2.566	
SIC	3.706		3.297		3.057		3.081		3.133		3.126		2.633		2.646	
HQC	3.690		3.276		3.032		3.044		3.124		3.110		2.589		2.597	
DW stat	0.709		1.488		1.979		1.975		1.753		1.797		2.067		2.067	
Q(1)	406.650	0.000	63.797	0.000	0.055	0.815	0.094	0.759	13.708	0.000	9.183	0.002	1.178	0.278	1.194	0.274
Q(2)	493.780	0.000	165.840	0.000	48.567	0.000	49.027	0.000	28.068	0.000	23.481	0.000	1.836	0.399	1.923	0.382
Q(8)	511.140	0.000	187.350	0.000	199.950	0.000	203.940	0.000	200.800	0.000	192.460	0.000	9.275	0.320	9.563	0.297
$Q^2(1)$	190.490	0.000	50.234	0.000	15.782	0.000	15.138	0.000	9.891	0.002	10.805	0.001	15.288	0.000	15.588	0.000
$Q^2(2)$	260.010	0.000	128.940	0.000	53.348	0.000	54.319	0.000	34.048	0.000	36.765	0.000	27.807	0.000	28.184	0.000
$Q^2(8)$	379.070	0.000	215.140	0.000	213.090	0.000	217.270	0.000	208.420	0.000	220.140	0.000	110.710	0.000	109.640	0.000

Note: The emboldened coefficients are significantly different from zero at a 0.05 significance level. The probability of Q-statistics for residuals that is greater than 0.05 indicates the existence of autocorrelation in residuals while the probability of Q-statistics for squared residuals that is greater than 0.05 indicates the clustering volatility.

Table 8: The estimates of TGARCH(1,1,1) regressions of SET50 index returns

	SET50(t)			SET50(t-1,t)			SET50(t-1,t+1)	
Variable	Coef.	S.E.	Variable	Coef.	S.E.	Variable	Coef.	S.E.
Constant	1.743	1.054	Constant	1.162	1.118	Constant	-0.891	1.683
Monday(t)	-0.090	0.094	Monday(t)	-0.022	0.091	Monday(t)	-0.078	0.158
Friday(t)	0.022	0.052	Friday(t)	0.039	0.052	Friday(t)	0.045	0.060
Holiday(t)	-0.112	0.086	Holiday(t)	-0.103	0.084	Holiday(t)	0.097	0.153
January(t)	0.088	0.064	January(t)	0.079	0.065	January(t)	0.007	0.094
Foreign(t)	**-0.593**	0.129	Foreign(t)	**-0.555**	0.132	Foreign(t)	-0.047	0.227
Local individual(t)	**-4.900**	0.139	Local individual(t)	**-4.890**	0.137	Local individual(t)	-0.219	0.422
Ln(Trade value(t))	**0.152**	0.046	Ln(Trade value(t))	**0.175**	0.049	Ln(Trade value(t))	0.049	0.071
S&P global return(t)	**0.073**	0.011	S&P global return(t)	**0.073**	0.011	S&P global return(t)	**0.072**	0.018
VIX(t)	0.000	0.003	VIX(t)	-0.001	0.003	VIX(t)	-0.006	0.006
FX return(t)	**-0.205**	0.063	FX return(t)	**-0.211**	0.063	FX return(t)	**-0.309**	0.103
SET50(t-1)	**-0.182**	0.025	SET50(t-2,t-1)	**0.813**	0.025	SET50(t-2,t)	**0.967**	0.032
			SET50(t-2)	**-0.845**	0.036	SET50(t-2)	**-1.028**	0.046
Sentiment(t)	**0.889**	0.112	Sentiment(t)	**0.914**	0.110	Sentiment(t)	0.059	0.142
Sentiment(t-1)	-0.077	0.135	Sentiment(t-1)	**-0.267**	0.101	Sentiment(t-1)	0.119	0.136
Sentiment(t-2, t-1)	**-0.197**	0.076	Sentiment(t-3, t-2)	**-0.150**	0.047	Sentiment(t-3, t-2)	0.016	0.075
Constant	**0.007**	0.003	Constant	**0.007**	0.002	Constant	**0.008**	0.002
resid2(t-1)	0.005	0.010	resid2(t-1)	0.001	0.010	resid2(t-1)	0.004	0.009
resid2(t-1) x I(resid(t-1)<0)	**0.083**	0.016	resid2(t-1) x I(resid(t-1)<0)	**0.088**	0.016	resid2(t-1) x I(resid(t-1)<0)	**0.093**	0.016
h(t-1)	**0.935**	0.016	h(t-1)	**0.938**	0.013	h(t-1)	**0.941**	0.011
R-squared	0.535		R-squared	0.768		R-squared	0.678	
Adjusted R-squared	0.528		Adjusted R-squared	0.765		Adjusted R-squared	0.673	
S.E. of regression	0.602		S.E. of regression	0.602		S.E. of regression	0.872	
Log likelihood	-827.2		Log likelihood	-820.5		Log likelihood	-1142.6	
Akaike info criterion	1.736		Akaike info criterion	1.726		Akaike info criterion	2.390	
Schwarz criterion	1.772		Schwarz criterion	1.764		Schwarz criterion	2.428	
Q(1)	0.440	0.507	Q(1)	0.262	0.609	Q(1)	0.240	0.624
Q(2)	0.639	0.727	Q(2)	2.576	0.276	Q(2)	0.248	0.884
Q(8)	6.491	0.592	Q(8)	8.424	0.393	Q(8)	3.048	0.931
Q2(1)	0.106	0.744	Q2(1)	0.037	0.848	Q2(1)	0.413	0.520
Q2(2)	0.267	0.875	Q2(2)	0.101	0.951	Q2(2)	1.167	0.558
Q2(8)	1.019	0.998	Q2(8)	1.011	0.998	Q2(8)	6.611	0.579

Note: The emboldened coefficients are significantly different from zero at a 0.05 significance level. The probability of Q-statistics for residuals that is greater than 0.05 indicates the existence of autocorrelation in residuals while the probability of Q-statistics for squared residuals that is greater than 0.05 indicates the clustering volatility.

Chapter 5 Summary

This study attempts to introduce a new leading indicator, a market sentiment index, for the SET, not to neither predict the movement of the SET or SET50 index nor find the best set of determinants for the SET or SET50 index changes. The use of textual analysis, Big Data and application programming interface (API) allows us to obtain information from various sources, including SET database and Datastream, online news and posts (written in Thai language) relating to the companies listed in the SET50. A Big Data analytic process is employed to extract tonal words from Thai online social media which will then be used to construct a composite sentiment index reflecting the, on average, pessimism/optimism toward listed firms expressed in Thai online and social media. The daily data for the period 2015 – 2018 are employed in this study.

We construct a sentiment index using statistics obtained from a Bag-of-Word analysis and the market values of SET50 index constituents. As the constituent list is changed every half year, the combination of firms in the sentiment index is changed every six months while the weight for sentiment toward each firm changes daily. In addition, we have employed some regressions to prove that the sentiment index we introduce has significant relationships with the changes in SET50 index. We show that our market sentiment index can signal changes in SET50 index from day t to day t+3.

This study has some limitations. Firstly, the observed and collected daily news might not be available for all SET50 stocks as 1) there might not be any news related to some of SET50 firms every single day; 2) the news related to the SET50 firms might not contain the identified tonal words, which are included in our sentiment dictionary. Future research and development of sentiment dictionary could provide more input into the Bag-of-Word analysis. Secondly, in a single news article, the content might be related to more than one stock. The team has tried to solve this challenge by using different keywords to identify each particular stock. Also, the paragraphing and sectioning of the news article are used to alleviate this inherent limitation. The third problem is caused by multiple meanings and interpretation of Thai tonal words/language in financial analysis context. This is a common limitation of similar types of projects as one of the key challenges in sentiment analysis is to understand semantics in groups of words (Sul et al., 2017).

Deep Machine learning can be used to tackle the challenge. However, due to the limited time and scope of this project, the machine learning for this challenge is not fully developed. The development of the Deep Machine learning from this project can be further improved and used as a basis for more complex machine training for Thai text analysis in the financial context.

Bibliography

Baek, C. (2016), "Stock prices, dividends, earnings, and investor sentiment", Review of Quantitative Finance and Accounting, Vol. 47 No. 4, pp. 1043-1061.

Bajo, E. and Raimondo, C. (2017), "Media sentiment and IPO underpricing", Journal of Corporate Finance, Vol. 46, pp. 139-153.

Bollen J., Mao H., Zeng X. 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1-8.

Bukovina, J. 2016. Social media and capital markets: An Overview, Social and Behavioral Sciences 220, 70-78.

Chung, S.-L., Hung, C.-H. and Yeh, C.-Y. (2012), "When does investor sentiment predict stock returns?", Journal of Empirical Finance, Vol. 19 No. 2, pp. 217-240.

Davis, A.K., Piger J.M., Sedor, L.S. 2006. Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. Working paper, Federal Reserve Bank of St. Louis. Fama, E. F. 1965. Random walks in stock market prices. Financial Analysts Journal, 1249-1266.

Day, M. and Lee, C. (2016), "Deep learning for financial sentiment analysis on finance news providers", in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1127-1134.

Dergiades, T. (2012), "Do investors' sentiment dynamics affect stock returns? Evidence from the US economy", Economics Letters, Vol. 116 No. 3, pp. 404-407.

Fama, E. F. 1970. Efficient capital markets: A review of theory and empirical work. Journal of Finance 25, 383-417.

Feng, L. 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan.

Fraiberger, S. P., Lee, D., Puy, D. and Ranciere, R. 2018. Media Sentiment and International Asset Prices. IMF Working Paper. WP/18/274.

Globalstats Statcounter. 2018. Social Media Stats Thailand [Online]. Available: http://gs.statcounter.com/social-media-stats/all/thailand [Accessed 1 December 2018]

Guo, K., Sun, Y. and Qian, X. (2017), "Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market", Physica A: Statistical Mechanics and its Applications, Vol. 469, pp. 390-396.

Joseph, K., Babajide Wintoki, M. and Zhang, Z. (2011), "Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search", International Journal of Forecasting, Vol. 27 No. 4, pp. 1116-1127.

Kim, S.-H. and Kim, D. (2014), "Investor sentiment from internet message postings and the predictability of stock returns", Journal of Economic Behavior & Organization, Vol. 107, pp. 708-729.

Lee, W. Y., Jiang, C. X. and Indro, D. C. (2002), "Stock market volatility, excess returns, and the role of investor sentiment", Journal of Banking & Finance, Vol. 26 No. 12, pp. 2277-2299.

Li, X., Xie, H., Chen, L., Wang, J., Deng, X. 2014. News impact on stock price return via sentiment analysis. Knowledge-Based Systems 69, 14-23.

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y., Ngo, D.C.L. 2014. Text mining for market prediction: A systematic review. Expert Systems with Applications 41, 7653-7670.

Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Systems Journal, 13(1), 760-770.

Ryu, D., Kim, H. and Yang, H. (2017), "Investor sentiment, trading behavior and stock returns", Applied Economics Letters, Vol. 24 No. 12, pp. 826-830.

Sayim, M. and Rahman, H. (2015), "The relationship between individual investor sentiment, stock return and volatility", International Journal of Emerging Markets, Vol. 10 No. 3, pp. 504-520.

Statista. 2018. Most famous social network sites worldwide as of October 2018 [Online]. Available:https://www.statista.com/statistics/272014/global-social-networks-ranked-bynumber-of-users/ [Accessed 1 December 2018]

Sul, H. K., Dennis, A. R. and Yuan, L. (2017), "Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns", Decision Sciences, Vol. 48 No. 3, pp. 454-488.

Sun, L., Najand, M. and Shen, J. (2016), "Stock return predictability and investor sentiment: A high-frequency perspective", Journal of Banking & Finance, Vol. 73, pp. 147-164.

Tetlock, P.C. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62(3), 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance 63(3), 1437-1467.

The Stock Exchange of Thailand. 2018. Investor Type:Yearly Cumulative since 1 Jan - 25 Dec 2018 [Online]. Available:https://marketdata.set.or.th/mkt/investortype.do [Accessed 25 December 2018]

Urquhart, A., Hudson, R. 2013. Efficient or adaptive markets? Evidence from major stock markets using very long run historic data. International Review of Financial Analysis 28, 130-142.

Verma, R. and Verma, P. (2007), "Noise trading and stock market volatility", Journal of Multinational Financial Management, Vol. 17 No. 3, pp. 231-243.

We are social. 2018. Digital in 2018: World's Internet Users Pass The 4 Billion Mark [Online]. Available:https://wearesocial.com/blog/2018/01/global-digital-report-2018 [Accessed 1 December 2018]

Yang, C. and Zhang, R. (2014), "Does mixed-frequency investor sentiment impact stock returns? Based on the empirical study of MIDAS regression model", Applied Economics, Vol. 46 No. 9, pp. 966-972.

Yang, H., Ryu, D. and Ryu, D. (2017), "Investor sentiment, asset returns and firm characteristics: Evidence from the Korean Stock Market", Investment Analysts Journal, Vol. 46 No. 2, pp. 132-147.

Yu, Y., Duan, W. and Cao, Q. (2013), "The impact of social and conventional media on firm equity value: A sentiment analysis approach", Decision Support Systems, Vol. 55 No. 4, pp. 919-926.

Zhang, X., Fuehres, H., Gloor, P.A. (2014). Predicting Stock Market Indicators through Twitter "I hope it is not as bad as I fear". Procedia - Social and Behavioral Sciences 26, 55-62.