ANALYZING SOCIAL MEDIA CONTENT TO GAIN COMPETITIVE INTELLIGENCE

Jitrlada Rojratanavijit

A Dissertation Submitted in Partial
Fulfillment of the Requirements for the Degree of
Doctor of Philosophy (Applied Statistics)
School of Applied Statistics
National Institute of Development Administration
2017

ANALYZING SOCIAL MEDIA CONTENT TO GAIN COMPETITIVE INTELLIGENCE

Jitrlada Rojratanavijit

School of Applied Statistics

Assistant Professor...Major Advisor
(Preecha Vichitthamaros, Ph.D.)

The Examining Committee Approved This Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Applied Statistics).

Associate Professor...Committee Chairperson
(Surapong Auwatanamongkol, Ph.D.)

Associate Professor...Committee
(Ohm Sornil, Ph.D.)

Assistant Professor...Committee
(Preecha Vichitthamaros, Ph.D.)

Assistant Professor..................S. Phongsuphap...............................Committee
(Sukanya Phongsuphap, Ph.D.)

Assistant Professor...Dean
(Sutep Tongngam, Ph.D.)

August 2017

ABSTRACT

Title of Dissertation	Analyzing Social Media Content to Gain Competitive Intelligence
Author	Miss Jitrlada Rojratanavijit
Degree	Doctor of Philosophy (Applied Statistics)
Year	2017

The emergence of social media in Thailand has given millions of users a platform to express and share their opinions about products and services, among other subjects, and so Twitter is considered to be a rich source of information for companies to understand their customers by extracting and analyzing sentiment from Tweets. The main goal was to investigate the possible uses of Twitter information for businesses in Thailand to take advantage of and to solve any associated limitations caused by the semantics of the Thai language. Hence, social media content, specifically Tweets were utilized to generate Competitive Intelligence (CI). A new method for Twitter sentiment analysis called ASTS was proposed by using both supervised learning and lexicon-based techniques. Experiments were conducted using Tweets on three mobile network operator companies: AIS, DTAC, and TRUEMOVEH obtained using the Twitter search API focused on Tweets in Thai. A total of 72,661 were collected over a period of six months (from October 1, 2014 to March 31, 2015).

ASTS consists of three modules: (1) data collection, (2) data pre-processing, and (3) classification and evaluation. A collection program was developed to search for keywords in the Twitter feed using the Twitter Search API and setting the language parameter “lang=th” and excluding reTweets. The process for the data pre-processing module was divided into three steps: (1) Text extraction from the Tweets, (2) Text pre-processing, and (3) Thai word segmentation. For the classification and evaluation module, the main intention was to identify opinion polarity, positive, negative, and neutral. The classification process was divided into two sub-modules: opinion filtering using supervised learning techniques and opinion polarity identification using lexicon-based techniques. Experimental results showed that the proposed method overcomes previous limitations from other studies and was very effective in most cases. The average accuracy is 84.80% with 82.42% precision, 83.88% recall and 82.97% F-measure. In particular, this clearly shows that opinion filtering helped to analyze Tweets more accurately.

A case study approach for CI in social media aptly demonstrated the use of ASTS. Out of a total of 20,269 Tweets, 9,631 mentioned AIS (47.52%), 7,099 mentioned DTAC (35.02%), and 3,539 mentioned TRUEMOVEH (17.46%). The sentiment scores from the analysis results of using ASTS showed the overall customer sentiment for the companies. The sentiment score for TRUEMOVEH (-0.27) was slightly better than AIS (-0.38) and DTAC (-0.45). Benchmarking against competitors is essential information for CI. Strength and weakness analysis on the companies was derived using radar charts of the benchmarking of sentiment scores on the top five keyword mentions (net, wifi, promotion, switching, and employee). Furthermore, examples of using CI in terms of monitoring, opportunity events, and early warning alerts were presented. Opportunity events can be advantageous in response to negative sentiment Tweets on competing companies and can help a company to entice customers away from competing companies. Early warning alerts are based on negative sentiment Tweets on a company from which it should quickly identify customer dissatisfaction and then correct the associated problem.

The results of this study show the usefulness of the proposed method for theoretical reference and as a practical guide. The findings from this analysis prove that CI extracted from social media content can help businesses to understand their customers' opinions and compare them with those of their competitors. As a result, this research illustrates that CI from analyzing social media content has great potential to produce useful information, actionable knowledge, and critical insights for companies to enhance competitiveness and solve business problems

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the help and support of several people and my organization. First, I would like to express my sincere gratitude to my advisor, Assistant Professor Dr. Preecha Vichitthamaros, for his continual support during my Ph.D study and related research, and especially for his patience, motivation, and immense knowledge.

I am sincerely grateful to Assistant Professor Dr. Sukanya Phongsuphap, Faculty of Information and Communication Technology, Mahidol University, who gave her valuable time to advise and train me as well as moderate this research and, along those lines, helped to significantly improve this manuscript.

I also have to thank the members of my PhD committee, Associate Professor Dr. Surapong Auwatanamongkol and Associate Professor Dr. Ohm Sornil, for their helpful comments and suggestions in general. In addition, I would like to express my gratitude to Associate Professor Dr. Duanpen Teerawanviwat, and Dr. Arnond Sakworawich for their invaluable lectures and helpful comments.

I would also like to extend my thanks to all of the staff at the School of Applied Statistics and the library of NIDA for their kind support. In addition, I am grateful to all my friends and other persons for their help, encouragement, and sharing of their knowledge and experience.

I am also immensely grateful to MEA for giving me the chance to study for a PhD by supporting me both financially and allocating me sufficient time to complete my studies.

Finally, I would like to thank my family for all their love and encouragement, especially my parents who raised me with love and have supported me in all my pursuits.

Jitrlada Rojratanavijit
August 2017

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

Figures **Page**

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

With the increasing usage of social media, they have undoubtedly become even more important as a source of qualitative and quantitative information and have become a most vital source of news and people's opinions on a wide variety of topics. Furthermore, social media have become important for communication and exchange of information and comprise a potent feedback channel for companies to understand consumers, especially the huge amount of the user-generated content steadily increasing on social networking sites (Zinner and Zhou, 2011). The emergence of social media tools has created a diversity of textual information containing hidden knowledge for businesses to leverage. In addition, the large amount of feedback on social media coming directly from customers has become a new source from which to mine what is referred to as competitive intelligence. In particular, marketers are able to sift through huge amounts of social media data to discover brand popularity and patterns of interest so to achieve a competitive advantage for companies over their competitors (He, Zha and Li, 2013). While the amount of text data explodes, text mining is a way to deal with it and is a helpful tool for companies to gain insight on their customers from social media content, which are a rich source of textual information. Applications that leverage unstructured data from online public communications to support marketing intelligence and business intelligence are divided in three categories: early alerting, buzz tracking, and sentiment mining (Glance, Hurst, Kigam, Siegler, Stockton and Tomokiyo, 2005).

In the face of ferocious market competition, there is a need for organizations to focus on time and speed, and in addition, quality and innovation (Dai, 2013: 2). Competitive intelligence (CI) is an essential component for organizations to gain a competitive advantage and leverage. Organizations use CI to compare themselves to other organizations (competitive benchmarking) to identify risks and opportunities in their markets, and to utilize this in business decision making. A recent study by Fan and Gordon (2014) mentioned that the growth of users on social media sites has increased steadily, so businesses need to monitor and use them to their advantage. In addition, they have described CI from social media analytics as being supportive to organizations by helping them understand their customers, suppliers, competitors, environments, and overall business trends.

Sentiment mining on social media content has become increasingly popular. Researchers from diverse fields have analyzed social media content to generate specific knowledge for their respective subject domains. For example, Gaffney (2010) analyzed Tweets with hashtag #iranElection using histograms, user networks, and frequencies of top keywords to quantify online activism. A similar study has been conducted on a natural disaster event: Hurricane Sandy; Dong (2013) explored the causality correlation between the approaching hurricane and the sentiment of the public toward it. Other research used to gauge business real world outcomes such as competitive analysis have been carried out in the pizza industry (He et al., 2013), predicting box-office revenues (Asur and Huberman, 2010), and analyzing business performance (Paniagua and Sapena, 2014).

Twitter, one of the most popular social media tools, claims that it has more than 550 million clients, out of which more than 271 million are dynamic (Wikipedia, 2014a). Twitter allows people to broadcast and share short (140 characters) real-time messages called Tweets which correspond to thoughts or ideas. Many people use it to send updates about their activities, as a tool for conversation, and to share information and report news (Java, Song, Finin and Tseng, 2007). Tweets may include one or more entities and reference places in their content, such as user mentions (@), hashtags (#), URLs, and media that may be associated with a tweet, and places are locations in the real world that may be attached to a Tweet (Wikipedia, 2014a).

Twitter has also become popular in Thailand with 4.5 million users, which places the country 17th in the world of global Twitter users (Saiyai Sakawee, 2014). The Internet and mobile technologies have been the main factor behind the growth of social media by providing technological platforms for information broadcasting, content generation, and two-way communication (Zeng, Chen, Lusch and Li, 2010). Moreover, Thailand mobile operator companies also offer broadband technologies such as 3G and 4G as well as high-speed downlink packet access at a cheap rate. As a result, Tweets have rapidly become a goldmine of information for companies to monitor their brands and more readily understand their customers.

1.2 Objectives of the Study

Social media networks such as Twitter and Facebook have become an important channel for communication throughout society. As mentioned above, social media analytics can provide "competitive intelligence" by helping businesses understand their customers, environments, suppliers, competitors, and overall business trends. The objectives of the study are:

1) To propose a framework/process for analyzing data from social media content, especially Twitter.

2) To study the possibility of Twitter data mining to gain CI using a case study of mobile network operators in Thailand.

1.3 Scope and Limitations

In this research, a case study was conducted to collect and analyze public social media data, only Twitter, to gain CI under the following scope:

1) In order to collect Tweets, the Twitter Search API was used and configured to extract only Thai language Tweets by setting the language parameter “lang=th” and excluding reTweets.

2) A case study was carried out using Tweets about the three largest mobile network operators in Thailand: AIS, DTAC, and TRUEMOVEH over a period of six months from October 1, 2014 to March 31, 2015.

3) Tweets that contained more than one brand were excluded because the message may include a comparative sentence, which is out of the scope of this study.

1.4 Expected Benefits of the Study

This research’s main contribution is to demonstrate CI via the mining and analysis of social media data, Twitter. It should be considered as a mixed concept of technical research and the application of computing techniques to business problems, and resulted in the following major outcomes:

1) The main goal is to investigate the possibility of using of Twitter information for businesses in Thailand to gain competitive intelligence.

2) The development and evaluation of the Acquisition of Sentiment from the Twitter System (ASTS), which fulfils the requirements of solving the research problem.

3) An illustration of how organizations can leverage social media analytics to enhance business value through a case study by analyzing public social media data of three competing companies.

1.5 Organization of this Dissertation

The rest of this dissertation is organized as follows. Chapter 2 provides a literature review as background for the study, and Chapter 3 contains the methodology for developing the tools and their evaluation. In addition, for comparing the performance of the proposed method with the original concept, the Twitter Opinion Mining framework (TOM) is also included. Chapter 4 contains an implementation of the proposed method with the real-life social media data from three companies in the telecommunication industry in Thailand. Finally, Chapter 5 provides conclusions on the study and raises future research motivated by this study.

CHAPTER 2

LITERATURE REVIEW

In this chapter, relevant literature is analyzed to offer the necessary background information and terminology as the foundation for the rest of this dissertation. The keywords used for searching the literature include: social media, Twitter, text mining, opinion mining, sentiment classification, competitive intelligence, SWOT analysis, customer insight, and related works.

2.1 Social Media

The term “social media” can be defined as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content” (Kaplan and Haenlein, 2010). With increasing usage of social media, it will undoubtedly become even more important as a source of qualitative and quantitative information. Social media has become the most vital source of news and people’s opinions on a variety of topics. The huge number of the user-generated content (social media content) has increased on the Internet because web–enabled devices such as cellphones and tablets are facilitating more people to access and create social media content with no geographic or economic boundaries. Albarran (2013: 4) classified the broader social media industries using several sub-markets of activity:

1) Social networking. Popular sites such as Facebook, Myspace, LiveJournal, and Tagged allow users to create personal profiles to share with others on their own networks.

2) Community/microblogging. Twitter, Kaboodle, and Fark are popular applications while Pinterest has been called “Twitter with pictures”. These are tools for sharing content.

3) Professional networking sites. These include Linkedin and Google+ which target the business and professional community. Slideshare is a related site which allows the sharing of presentations.

4) Social tagging. StumbleUpon, Del.icio.us, and Technorati are the most well-known sites in this area.

5) Image/Photo sites. Pinterest, Flickr, and Instagram are renowned sites for sharing photos and content.

6) Video sites. These include YouTube, Vimeo, and Socialcam.

7) Social news. Digg, Reddit, Newsvine, and Yahoo Buzz are the most popular sites.

8) Gaming sites. Zynga is the market leader in this area. Other firms involved in gaming include Pogo, Yahoo! Games, Big Fish, and PopCap.

9) Consumer shopping. Groupon, Living Social, Dealster, and Social Buy are the popular sites which offer daily digital coupons or special offers.

10) Review sites. Yelp, Citysearch, and Epinions offer customer reviews for others to share.

11) Wikis. Wikipedia is the best-known site that allows the sharing and editing of content posted online. Investopedia, Wikitravel, Wikia, and ShopWiki are also popular sites.

12) Social publishing. Scribd is a popular publishing site.

13) Location-based services (LBS). Foursquare was the first to capture wide attention and is often used in combination with review and shopping sites.

2.2 Twitter

A popular social media tool for sharing content is Twitter. In 2014, Twitter had more than 550 million clients, out of which more than 271 million were dynamic clients (Wikipedia, 2014a). Many people use it to update their activities, as a conversation tool, and to share information and report news (Java et al., 2007). This tool has generated a wealth of textual data which contain hidden knowledge for businesses to leverage to obtain a competitive edge. Hence, it would be very useful and challenging for a company to transform the huge amount of unstructured, ungrammatical, and noisy data from this source into useful information.

Paniagua et al. (2014) mentioned that Twitter is a free microblogging service via which interactions are normally based on mutual affinity, while conversely, Facebook is a complex networking tool via which relationships are constructed based on friendship or acquaintance. Moreover, they also infer that Twitter has become an interesting tool that constitutes a valuable resource due to the specific characteristics of microblogging, such as ambient awareness and a push-and-pull communication format.

Twitter users first create a profile page that includes a photo, location, website address, and a short biographical sketch (Goff, 2013). Twitter allows people to broadcast and share short (140 characters) real-time messages called *Tweets* which correspond to thoughts or ideas (Wikipedia, 2014a; Russell, 2013). Twitter users can reply to Tweets or mention them in Tweets to others and also forward or “reTweet” (RT) interesting Tweets to others. Furthermore, Twitter enables users to find and follow the Tweets of others and to also search for Tweets related to key words or topics (Goff, 2013).

Tweets may include one or more entities and reference places in their content, such as user mentions, hashtags (#), URLs, and media that may be associated with a Tweet, and places are locations in the real world that may be attached to a Tweet (Russell, 2013). Let’s consider a sample Tweet with the following text: “@rjitrlada is writing an article from her office in Bangkok. #socialmedia : http://bit.ly/1utr1aR”. The Tweet is 97 characters long and contains three Tweet entities: the user mention @rjitrlada, the hashtag #socialmedia, and the URL http://bit.ly/1utr1aR. Furthermore, users can use emoticons (punctuation and letters to express their feelings), a target (specific to someone by using the “@” symbol), and hashtags (to mark topics with keywords by using the “#” symbol) (Agarwal, Xie, Vovsha, Rambow and Passonneau, 2011). Besides, Twitter has the characteristics of a social network; an information-sharing network whereby both user-generated content and regular media are propagated publicly although relationships do not need to be connected (Mogollón, 2014). Java et al. (2007) found that the main intentions of Twitter users are as described below:

1) Daily chatter: the most common use of Twitter is talking about daily routines or what you are currently doing.

2) Conversations: users use Twitter to reply to or send direct messages by using the @ symbol followed by a username.

3) Sharing information: users send messages that contain URLs to share more information. They almost always use a URL shortening service like TinyURL to make this feature feasible.

4) Reporting news: users use Twitter to report latest news or comment about current events. They can update weather reports and new stories from RSS feeds using developer APIs.

For a Twitter-based stakeholder communication strategy for firms, Tran (2012) assessed the intention to use Twitter based on five main stakeholder groups as follows:

1) Consumer-focused: Tweets provide information about products, promotion, new locations and news or events. Besides, Tweets that reply to customer's questions or complaints or comment are included too.

2) Investor-focused: Tweets always mention company's business plan such as the opening of new offices, new product development projects, executives, and mergers. In addition, they may report a company's revenue, stock price, market share, and advantages over its competitors.

3) Employee-focused: Tweets enable information sharing and communication within a company. Such Tweets may cover subjects such as company-related updates and interaction between employees to support customers.

4) Government-focused: Tweets frequently mention issues such as jobs, taxes, and security. They may cover how government policies have an effect on a particular company's business and how a company's business provides value to the nation.

5) Community-focused: Tweets provide information about company activities or company-sponsored community activities to support the environment, education, health, children, and charities.

In Thailand, Saiyai Sakawee (2014) reported that Zocial Inc, which monitors social media trends in the Thailand, summarized data and posited that Twitter is now popular in the country. In 2014, there were 4.5 million Twitter users, which is a 350% growth from the previous year, as shown in Figure 2.1. Pawoot Pongvitayapanu (2014) mentioned that among Twitter users globally, Thailand ranked 17th in the world. He presented that Thai people send 5 million Tweets per day, of which 1.9 million are reTweets. Besides, he also mentioned that the number of average active Twitter users is 250,000 per day and they prefer to use the app more often between 9 pm and 10 pm.

Zinner et al. (2011: 71) mentioned that “Social media feedback is usually more emotional – that it strongly slanted toward and opinion.” In comparison, traditional feedback from customers via other channels typically contains less emotion and feelings. Thus, social media feedback could help companies to consider how customers really feel about an issue. However, the data from social media are unpredictable and the amount is growing exponentially, and so machine learning and text mining are applied to the vast amount of social media data to detect and discover new knowledge and interesting patterns. This information could be helpful for companies to understand what customers really want and what their competitors are doing (Dey, Haque, Khurdiya and Shroff, 2011).

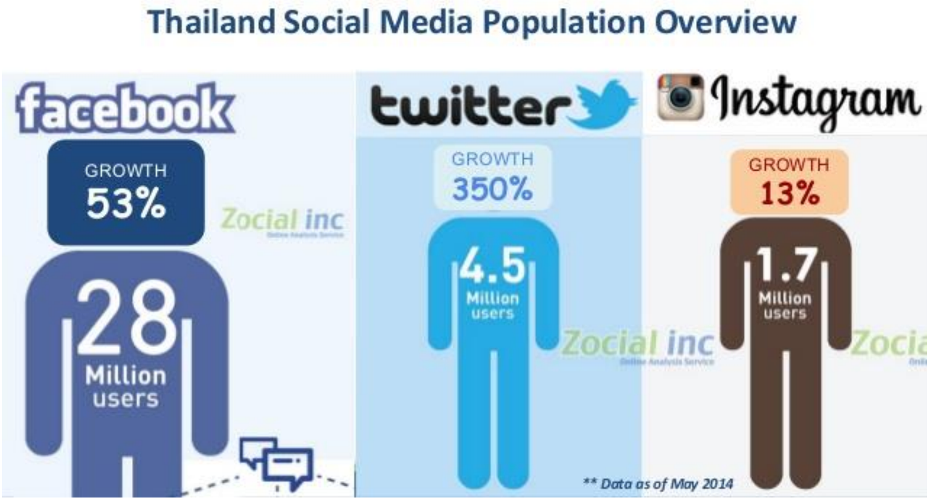


Figure 2.1 Social media overview for Thailand in 2014

Source: Pawoot Pongvitayapanu, 2014: 5.

2.3 Text Mining

While the amount of text generated in electronic form is exploding, text mining is a way to deal with this flood of text data. The vast amount of textual information on the Internet has undoubtedly fueled a rapid rise in text mining (Davenport and Harris,

2007) being applied to extract new trends or relationships from the overwhelmingly large amount of digital data. Its main purpose is to automatically identify hidden knowledge in unstructured text, and then create interpretation or models that explain interesting patterns and trends in the textual data (He et al., 2013), which is probably the reason why text mining has captured significant attention in the business and research communities concerned with its practical application. All text mining applications deal with solutions to meet that challenge in one form or another. Miner, Elder, Fast, Hill, Nisbet and Delen (2012) defined five basic types of analytical text mining applications and use cases as follows:

1) Extracting "meaning" from unstructured text. This application involves the understanding of core themes and relevant messages in a corpus of text without actually reading the documents.

2) Automatic text categorization. Automatically classifying text is an efficient way to organize it for downstream processing.

3) Improving predictive accuracy in predictive modeling or unsupervised learning. Combining unstructured text with structured numeric information in predictive modeling or unsupervised learning (clustering) is a powerful method to achieve better accuracy.

4) Identifying specific or similar/relevant documents. Efficiently extracting those documents from a large corpus of text that are relevant to a particular topic of interest or are similar to a target document (or documents) is a vitally necessary operation in information retrieval.

5) Extracting specific information from the text ("entity extraction"). Automatically extracting specific information from the text (such as names, geographical locations, and dates) is an efficient method for presenting highly focused information for downstream analytical processing or for direct use by decision makers.

Moreover, text mining is helpful for a company to gain customer insight. Glance et al. (2005) stated that applications that leverage data from online public communications to support marketing intelligence and business intelligence are divided into three categories as follows:

1) Early alerting – informing subscribers when a rare but critical, or even fatal, condition occurs.

2) Buzz tracking – following trends in topics being discussed and understanding what new topics are forming.

3) Sentiment mining – analyzing people's opinions, sentiments, evaluations, attitudes, and emotions toward entities such as products, services, and organizations (Miner, Elder, Fast, Hill, Nisbet and Delen, 2012; Liu, 2012; Wilas Chamlertwat, 2011; Alisa Kongthon, Choochart Haruechaiyasak, Chatchawal Sangkeettrakarn, Pornpimon Palingoon and Warunya Wunnasri, 2011).

2.4 Opinion Mining

Opinion mining, also known as sentiment analysis, is the process of recognizing those opinions that a particular discourse expresses. It attempts to automatically identify people's opinions from a text written in their natural language (Dai, 2013: 30). There are several ways to classify the sentiment present in a document with one of them being its polarity as positive, negative, or neutral. According to Liu (2012), sentiment analysis is the field of study that analyzes human opinion, sentiment, evaluation, appraisal, attitude, and emotion toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Sentiment analysis can be categorized according to the granularity of text and previous work has mainly focused on three levels (Liu, 2012):

1) Document/text level

The analysis at this level is to determine whether sentiment expressed in a whole document is positive or negative.

2) Sentence level

The analysis at this level is to determine whether the opinions expressed in a sentence is positive, negative, or neutral.

3) Aspect / feature /entity level

Unlike document or sentence-level analysis, aspect-level analysis explores what the holder likes or hates about the target.

Besides, research can also be conducted at phrase, clause, or word level. Sentiment analysis of Tweets is considered to be at document level (Yuan, 2016). However, since Twitter allows a maximum of 140 characters, each Tweet tends to be very short. Usually a Tweet contains only one simple sentence or just several words. Therefore, Twitter sentiment analysis also calls for a wide variety of strategies utilized for other levels of analysis.

2.5 Sentiment Classification

The methods for classification sentiment can be generally categorized into two methods: lexicon-based and machine learning-based.

2.5.1 Lexicon-based Methods

The lexicon-based approach relies on a sentiment lexicon (a collection of known and precompiled sentiment terms). It is made up of a list of words or phrases that convey positive or negative polarity information (Liu, 2012) and is a very important resource in sentiment analysis which is divided into a dictionary-based approach and a corpus-based approach using statistical or semantic methods to find sentiment polarity. The dictionary-based approach depends on finding opinion seed words, and then searches the dictionary for their synonyms and antonyms, and automatic expansion explores pair-wise word relations and generates a lexicon of proper size. Although dictionary-based approaches can generate a large number of sentiment words, those words are usually context and domain independent. Corpus-based approaches begin with a seed list of opinion words, and then find other opinion words in a large corpus to help in finding opinion words with context-specific orientations.

2.5.2 Machine Learning-based Methods

Text classification methods using the machine learning approach are roughly divided into supervised and unsupervised learning. The supervised methods make use of a large number of labeled training documents (the training dataset). The process of supervised learning algorithms are from the training dataset that can be thought of as a teacher supervising the learning process. The correct answers will be defined and the algorithm iteratively makes predictions on the training data that may require correction by the teacher. There are many kinds of supervised classifiers with the most popular being Naïve Bayes (NB) and Support Vector Machines (SVMs). The NB classifier is the simplest and most commonly used classifier and computes the posterior probability of a class based on the distribution of the words in the document (Bifet and Frank, 2010). For SVM classifiers, they are used to determine linear separators in the search space which can best separate the different classes (Liu, 2012).

Unsupervised methods are usually used when it is difficult to identify any labeled training documents. Furthermore, there is no external teacher or critic to oversee the learning process. In other words, there are no specific examples of the function to be learned by the system. Rather, provision is made for a task-independent measure of the quality or representation that the system is required to learn. That is to say, the system learns statistical regularities of the input data and develops the ability to learn the feature of the input data, thereby creating new classes automatically.

2.6 Competitive Intelligence

These days, there are many definitions of CI which can be divided into two categories (Dai, 2013: 21). One definition looks at CI as a kind of meaningful information that enables an organization to be aware of its external competitive environment through a continuous systematic collection process. The other definition sees CI as a process of information analysis and a vital component of a company's strategic planning and management process. The Strategic of Competitive Intelligence Professionals (SCIP) defines CI as a systematic and ethical process for gathering, considering, and analyzing market and external information that can affect a company's plans, decisions, and operations (SCIP, 1986; Wikipedia, 2016).

Dey et al. (2011) broadly classified CI into two categories of information depending on whether it is used for long-term or short-term planning. Strategic intelligence focuses on long-term issues and is used to analyze a company's competitiveness over a specified period into the future, while tactical intelligence concentrates on providing information that can influence short-term decisions. Regularly, this has been identified with investigation of a company's current market share and competitive landscape. Tarraf (2005) studied CI and concluded that competitor analysis is a crucial component of the CI system which includes data gathering, sorting and cleaning, analyzing, and presenting the results to decision makers.

Dey et al. (2011) classified tactical market intelligence into four categories as follows:

1) Market information – provides information about the popularity of competitors in terms of their products or brands as a whole, which products are moving in the market, and the market share of competitors. The sentiment of consumers related to the organization and its competitors also belong in this category.

2) Price information – provides knowledge about the prices of competitors' products.

3) Promotion – provides information about promotion strategies and the kind of promotional activities that are adopted by competitors.

4) Other issues – organizational information about competitors such as their workforce structure, internal shift in focus or vision, success or failure of their trials, new product launches, technology investments, etc., all of which contribute toward building profiles on competitors that can be useful to organizations.

Organizations use CI to compare themselves to other organizations (competitive benchmarking) so as to identify risks and opportunities in their markets. The most used method is SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis to realize the importance of knowing what their competitors are doing and how the industry is changing, and the information gathered allows organizations to understand their strengths and weaknesses (Wikipedia, 2016).

2.7 SWOT Analysis

A SWOT analysis summarizes the key issues from the business environment to discover threats and opportunities. It also evaluates the strengths and weakness of a company in relation to its competitors to identify the issues that it is facing or will face (Dai, 2013: 35). Strengths and weakness are factors in the internal environment of a company compared with the capability of competitors concerning, for example, technology, personnel, products, markets, and management structure. Strengths (S) are characteristics of a company that give it an advantage over others, weaknesses (W) are characteristics of a company that place it at a disadvantage relative to others, opportunities (O) are elements in the environment that a business or project could exploit to its advantage, and threats (T) are elements in the environment that could cause trouble for a business or project (Wikipedia, 2015b).

The SWOT analysis is a basic powerful strategic analysis tool. It causes one to strategically think about past, present, and future issues that may affect planning. The SWOT analysis may view the internal factors as strengths or as weaknesses depending upon their effect on the organization's objectives. The factors may include all of the 4P's of marketing: Product, Price, Place, and Promotion as well as personnel, finance, manufacturing capabilities, and so on (Wikipedia, 2015b). Additionally, the results of SWOT analysis will help the company as follows:

1) To explore new solutions or product with using strengths

2) To minimize the weaknesses

3) To leverage the opportunities

4) To revise plans to avoid threats

2.8 Customer Insight

Business and organizations need consumer insight into making better business decisions in marketing or services. In the psychological context, consumer insight is knowing what consumers think and feel about a company's products, services, or brand (Merlin, Bond and Foss, 2004). In recent years, customers have become more demanding and their expectations have risen, which makes it challenging for companies to satisfy their desires and have an increasingly shorter time period in which to respond to them (Kumar and Reinartz, 2012). The first step toward gaining customer insight is listening to customer feedback. This phenomenon, otherwise known as voice of the customer (VOC) is a term used in business for describing the in-depth process of gathering a customer's expectations, preferences, and dislikes. The traditional approaches for obtaining VOC are research (both quantitative and qualitative) that produces a detailed set of customers' wants and needs (Wikipedia, 2014b). However, these methods are time consuming and expensive, and the weakness of traditional market research has now been made more obvious by the emergence of social media (Zinner et al., 2011).

Social media is a potent feedback channel for companies to understand consumers, especially with the huge amount of user-generated content still increasing on social network sites (Zinner et al., 2011). People share opinions online about products or services to inform themselves, then content is disseminated publicly because it is no longer limited to people they actually know. Hence, companies must understand this fundamental shift in customer behavior and properly respond to it.

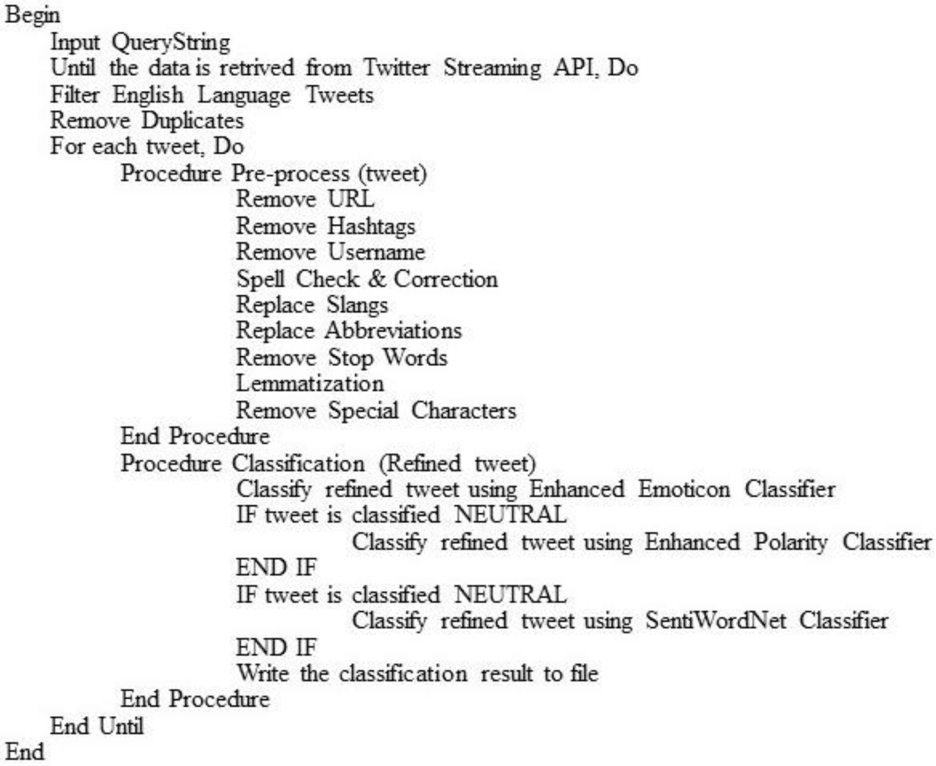
Kumar et al. (2012) recommended that the companies should take advantage of social media in at least three key areas. First, companies can get to know their customers better by listening to them online. Online content helps companies identify what customers really prefer or dislike. Following this, they can use this information to develop products, strategies, and success measures. Second, online customers expect to get an immediate response from companies, so the latter should improve their response time to customer feedback in order to meet their customers' needs. Third, companies can use social media to advance communication or marketing strategies to their customers.

2.9 Related Research

Researchers from diverse fields have analyzed social media content to generate specific knowledge for their respective subject domains. For example, Gaffney (2010) analyzed Tweets with hastag #iranElection using histograms, user networks, and frequencies of top keywords to quantify online activism. A similar study was conducted for a natural disaster event: Hurricane Sandy (Dong, 2013). He explored the causality correlation between an approaching hurricane and the sentiment of the public.

Khan, Bashir and Qamar (2014) proposed a new Twitter opinion mining (TOM) framework to categorize the polarity of Tweets into positive, negative, or neutral sentiment by applying a variant of techniques for Twitter feed analysis and classification which involved pre-processing steps and a hybrid scheme of classification algorithms. The proposed pre-processing steps included: removal of URLs, the hashtag symbol, usernames, and special characters; spelling correction using a dictionary;

substitution of abbreviations and slang with expansions; lemmatization; and stop words removal. They proposed a classification algorithm incorporating a hybrid scheme using emoticon analysis, an improved polarity classifier using a list of positive and negative words, and SentiWordNet analysis, as shown in Figure 2.2. In their research, the average accuracy of the TOM framework was 85.7% with 85.3% precision and 82.2% recall. However, the final results may have been contaminated by news and other information.



```
Begin
    Input QueryString
    Until the data is retrived from Twitter Streaming API, Do
    Filter English Language Tweets
    Remove Duplicates
    For each tweet, Do
          Procedure Pre-process (tweet)
                    Remove URL
                    Remove Hashtags
                    Remove Username
                    Spell Check & Correction
                    Replace Slangs
                    Replace Abbreviations
                    Remove Stop Words
                    Lemmatization
                    Remove Special Characters
          End Procedure
          Procedure Classification (Refined tweet)
                    Classify refined tweet using Enhanced Emoticon Classifier
                    IF tweet is classified NEUTRAL
                              Classify refined tweet using Enhanced Polarity Classifier
                    END IF
                    IF tweet is classified NEUTRAL
                              Classify refined tweet using SentiWordNet Classifier
                    END IF
                    Write the classification result to file
          End Procedure
    End Until
End
```

Figure 2.2 The polarity classification algorithm of the TOM framework

For the Thai language, Choochart Haruechaiyasak, Alisa Kongthon, Pornpimon Palingoon and Chatchawal Sangkeettrakarn (2010) proposed a framework for constructing a Thai language resource for feature-based opinion mining obtained from hotel reviews. They constructed a set of patterns from a tagged corpus, and then automatically extracted patterns and collected more sub-features and polar words from an untagged corpus. Later, Choochart Haruechaiyasak, Alisa Kongthon, Pornpimon Palingoon and Kanokorn Trakultaweekoon (2013) proposed S-Sense (a framework for analyzing sentiment from Thai social media content). They collected data from Twitter posts and the Pantip web board in the mobile service domain. Following this, they applied the Naïve Bayes algorithm to learn the classifier models. They manually labeled texts with appropriate intension and sentiment classes. Their lexicon consisted of general terms from dictionaries and clue terms which helped to identify the intension and sentiment. In an intension analysis experiment, they trained a binary classification model with two classes: related and others, to analyze four different intensions (announcement, request, question, and sentiment). For sentiment analysis, they trained a binary classification model with two classes: positive and negative with an accuracy of 91.64%. However, there is the possibility that content could be neutral, which was not considered in the study.

There are some researches that have used social media for business real-world outcomes. Dey et al. (2011) built a CI gathering and processing system that could be customized to collect different types of data from the web. Other researches have been used on business real-world outcomes, such as competitive analysis in the pizza industry (He et al., 2013), predicting box-office revenues (Asur et al., 2010), and analyzing business performance (Paniagua et al., 2014).

In 2015, a competitive analytics framework with sentiment benchmarks was proposed for industry-specific marketing intelligence (He, Wu, Yan, Akula and Shen, 2015). The procedures of the framework were divided into four steps. First, firms need to select a few leading companies in the industry and identify their social media sites for competitive analysis. Second, firms need to monitor the selected social media sites and constantly collect user-generated data posted on those sites. Third, a data pre-processing step is needed to transform raw data into a usable format. Next, a combination of various text mining, sentiment analysis, and traditional social network analysis techniques are used to examine the datasets to gain insights into users' social media activities and sentiments. Finally, the results of the social media analytics should be carefully reviewed and then used to derive insight, create intelligence reports, support decision making, and/or make recommendations.

He et al. (2015) proposed a framework to gain CI from social media data to enhance business value and market intelligence. They conducted a case study to collect and analyze a dataset with half a million Tweets related to two of the largest retail chains in the world: Walmart and Costco in the three months from December 1, 2014 to February 28, 2015. They mentioned that social media data can be divided into two main categories: social media data-at-rest (weekly or monthly analysis) and social media data-in-motion (real-time analysis). VOZIQ, a social media analytics tool, was used in this study. This tool was set up on the Amazon AWS cloud platform to collect Tweets that contained specific keywords utilizing the Twitter search API on a daily basis. Lexalytics, a popular sentiment analysis tool, was used to detect sentiment in each Tweet in their dataset. They used a well-known text analysis tool, NVivo 10, to query the product level from the collected Tweets, and then analyzed the content and sentiment of these products. In addition, Leximancer was used to mine and cluster the Tweets related to each of the products to gain a better understanding of what customers are talking about for each of them. The results of the case study revealed the value of analyzing social media mentions by conducting sentiment analysis and comparing it on the individual product level, and illustrated that a great deal of potentially useful information, actionable knowledge, and critical insights were obtained.

Kim, Dwivedi, Zhang and Jeong (2016) found that social media data contain CI by using an exploratory test with a multiple case study approach to compare two competing smartphone manufacturers: iPhone and Samsung in the past four months from August to December 2014. They conducted opinion mining and sentiment analysis, then followed up by further validation of results using statistical analysis. They employed a Twitter API, the TwitterR package, in the R software program to collect Tweets by using hashtag “#iPhone6” and “#GalaxyS5”. The program ran every day and the total number of Tweets in the dataset was 229,948. Because of their previous study showed that 90% of Tweets included irrelevant mentions (include reTweets). Therefore, the first step was to filter Tweets from marketing agencies and commercial accounts. Some web sites such as Twitterfeed, Hootsuit, and IFTTT were doubly publishing the same commercial Tweets. Therefore, the source field of a Tweet (URLs) was used to eliminate them.

They used the following three research questions to explore CI for competing companies in the same industry:

RQ1: Is there significant gap in social media volume between competitors with different market shares?

RQ2: Is there a significant gap discrepancy in levels of purchase intention between competing brands?

RQ3: Is there a significant gap in consumer sentiments about the two rivals?

The results showed that the volume of Tweets revealed a significant gap between the market leader and follower. The purchase intention data also reflected this gap, albeit to a lesser extent. In addition, social opinion was able to explain the sales performance gap between the competitors, and the researchers found that the social opinion gap was similar to the shipment gap.

Previous studies have shown that the social media space has become a common tool for communicating, networking, sharing content, and promoting brands. Companies can use social media as a customer service tool for listening to their customers and addressing their concerns. Social media can provide information about brand popularity, consumer sentiments, and competitor promotions. In addition, social media intelligence is a new term from social media analytics and its aim is to “derive actionable information from social media in context-rich application settings, develop corresponding decision-making or decision-aiding frameworks, and provide architectural designs and solution frameworks for existing and new applications that can benefits from the ‘wisdom of crowds’ through the Web.” (Zeng et al., 2010: 7). In addition, social media analytics also provide CI that helps businesses understand their environments, suppliers, competitors, and overall business trends (Fan et al., 2014). A recent literature review revealed that there are few studies to guide business for social media CI, so this is still a challenging area and could be quite useful for Thai businesses.

CHAPTER 3

METHODOLOGY

Methodology of this research is defined in two phases. First phase is the creating method for analyzing data from social media content, Twitter. For second phase, a framework/process is defined for gaining CI using a case study of mobile network operators in Thailand. Since the main goal is to solve the problem of extracting Twitter information for businesses in Thailand to take advantage of, a technique of sentiment analysis of Twitter data generated by Thais with the main focus being on Tweet classification and solving data sparsity issues is proposed. There are significant differences between written Thai and English. English has twenty-six letters, whereas Thai has forty-four consonant letters (Thai: พยัญชนะ, phayanchana), fifteen vowel symbols (Thai: สระ, sara), and four tone diacritics (Thai: วรรณยุกต์ or วรรณยุต, wannayuk or wannayut) (Wikipedia, 2015c). Besides, there is no punctuation or spaces between words in sentences in Thai, and so information must be segmented before feeding into the classifier. Therefore, existing text mining and sentiment analysis techniques cannot be directly applied to the Thai language.

To overcome these drawbacks, a number of challenges need to be overcome: classification accuracy, sarcasm, word usage variation, and data sparsity (Choochart Haruechaiyasak et al., 2013; Khan et al., 2014). The reason for these issues is variations in the use of slang words and other abbreviations because of the limitation of a Tweet (140 characters). The main idea is to pre-process the raw data and operate variant transformations to deal with slang, transliterated words, abbreviations, and other noise. For the classification process, a mixed method of supervised learning and lexicon-based techniques to filter Thai opinions and classify them into positive, negative, or neutral sentiment is proposed.

In the first phase, a Thai opinion mining method based on the techniques for Twitter feed analysis and classification is developed. The process is subdivided into three modules: (1) data collection, (2) data pre-processing, and (3) classification and evaluation. Tweets were obtained from the Twitter search API using query strings. Data pre-processing was used to extract the Tweets for text pre-processing and Thai word segmentation. For the classification and evaluation module, the main objective was to identify the polarity of Thai opinion Tweets. The proposed method is shown in Figure 3.1.

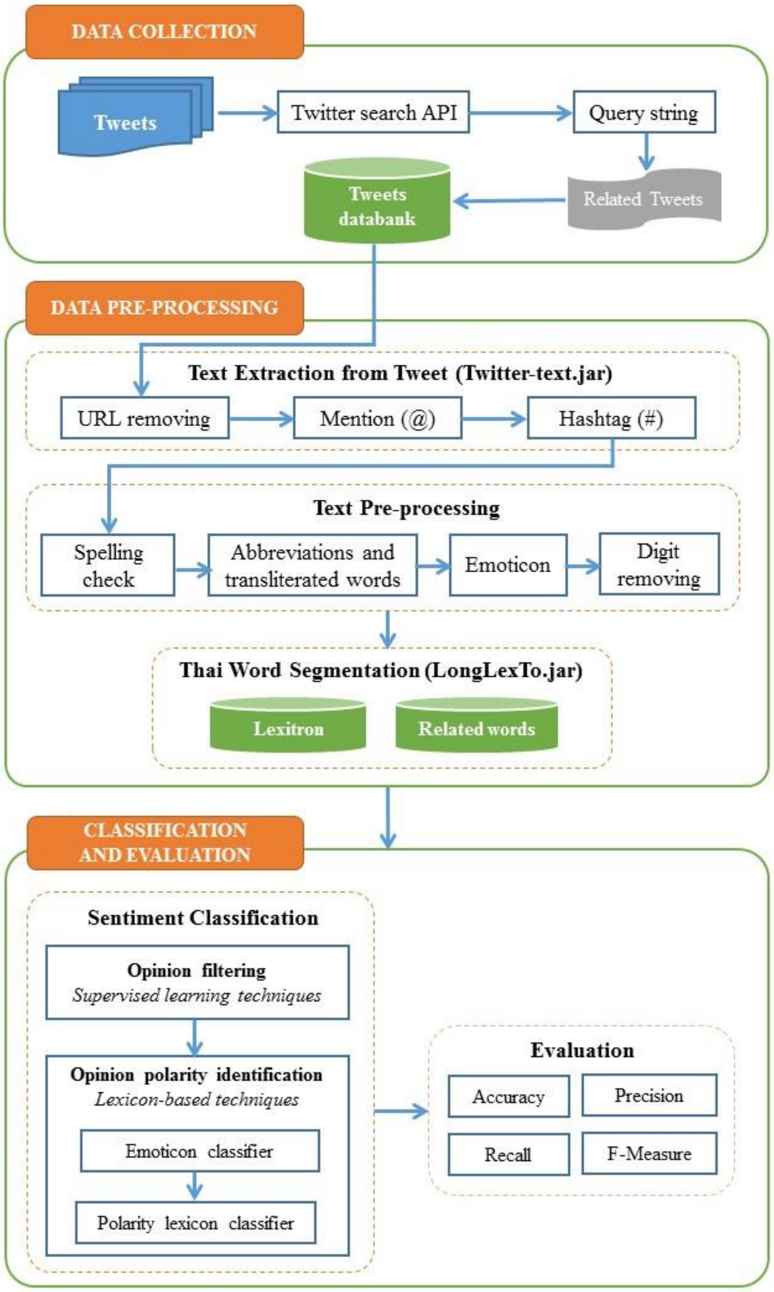


Figure 3.1 Acquiring Sentiment from the Twitter System (ASTS)

3.1 Context of the Study

The telecommunication industry in Thailand, particularly the mobile service market, is dominated by three major players, namely Advanced Info Service Public Company Limited (AIS), Total Access Communication Public Company Limited (DTAC), and True Move H Universal Communication Company Limited (TRUEMOVEH) in a ferociously competitive market. The National Broadcasting and Telecommunications Commission (NBTC), which has responsibilities for telecommunication services in Thailand, revealed that the number of mobile phone numbers in circulation is almost 100 million. As of June 2013, the market penetration rate was 131.84% over a population estimate at around 67.373 million, consisting of AIS (42 million), DTAC (30.6 million), and TRUEMOVEH (23.2 million). Figure 3.2 illustrates the market shares in the telecommunication industry in Thailand in 2015: AIS is the largest, DTAC is second, and TRUEMOVEH is third.

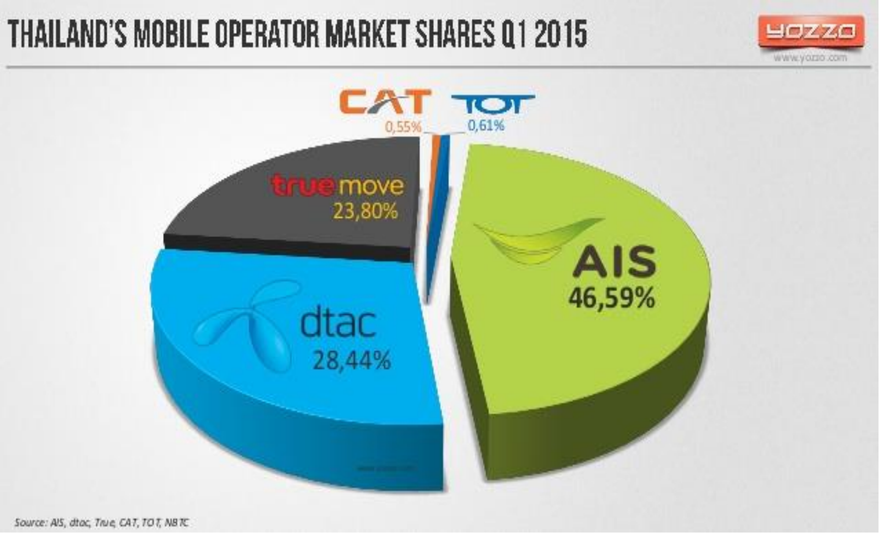


Figure 3.2 Thailand’s mobile operators’ market share Q1 2015
Source: Yozzo, 2015: 15.

In addition, this industry has seen more competition since NBTC announced the right of consumers to switch mobile operators without changing their mobile numbers since September 22, 2014. As a result, the three mobile operators have launched many marketing campaigns through various marketing channels such as TV advertising, newspapers, magazines, and direct mail. In addition, companies are increasingly using social media such as Twitter to promote products and services and to communicate with customers, (Figure 3.3). Their official Twitter accounts are ais_thailand (changed from ais_privilege since 20 November, 2014) for AIS, dtac for DTAC, and truemoveh for TRUEMOVEH.

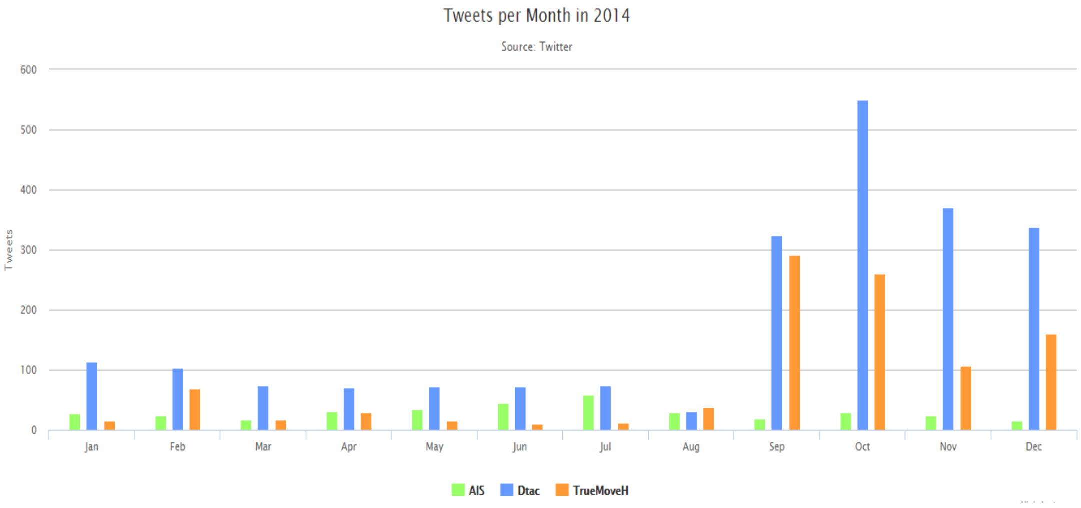


Figure 3.3 Posted Tweets from the three largest mobile operators in Thailand

3.2 Data Collection

To evaluate the proposed method in a real-world business context, a case study of the telecommunication industry in Thailand was conducted to collect and analyze a dataset of Tweets related to the three aforementioned largest mobile operator companies in Thailand: AIS, DTAC, and TRUEMOVEH in the past six months from October 1, 2014 – March 31, 2015. A collection program was developed to search for Tweets of interest from the Twitter feed using keywords and the Twitter search API, which allows queries against the indices of recent or popular Tweets. The API was configured to extract only Thai language Tweets by setting the language parameter “lang=th” and excluding reTweets. After this, the Tweets were kept in a databank that acted as input data for the pre-processing module.

Keywords related to three mobile network operators were used for searching, as shown in Table 3.1. As many social media sites have millions of users and the amount of data often grows at high speed, it is not always possible to save the data in full due to resource constraints. Thus, in this case study, 72,661 Tweets in total from the six months from October 1, 2014 to March 31, 2015 were captured as the dataset for analysis and kept in the Tweets databank, as shown in Table 3.2.

Table 3.1 Related keywords for mobile network operators in Thailand

Mobile Operator Company	**Keyword for Query**
1. Advanced Info Service Public Company Limited (AIS)	AIS, AIS_Privilege, AIS_Thailand,12Call, เอไอเอส
2. Total Access Communication Public Company Limited (DTAC)	DTAC, Trinet, ดีแทค
3. True Move H Universal Communication Company Limited (TRUEMOVEH)	TRUEMOVEH, TRUEMOVE, ทรูมูฟ
4. Others	3G, 4G, Edge, Wifi

Table 3.2 Structure of the table for storing Tweet information

Field	**Type**	**Description**
id	Integer	The integer representation of the unique identifier for this Tweet.
created_at	String	UTC time when this Tweet was created.
text	String	The actual UTF-8 text of the status update. See Twitter-text for details on what is currently considered valid characters.
source	String	Utility used to post the Tweet, as an HTML-formatted string.

Table 3.2 (Continued)

Field	Type	Description
retweet_count	Integer	Number of times this Tweet was reTweeted.
name	String	The name of the user as defined (not necessarily a person's name).
screen_name	String	The screen name, handle, or alias that users identify themselves with. screen_names are unique but subject to change.
user_created_at	String	The UTC datetime that the user account was created on Twitter.
favourites_count	Integer	The number of Tweets this user has favorited in the account's lifetime.
lang_user	String	Language identifier
location	String	The user-defined location for this account's profile.
profile_image_url	String	A HTTP-based URL pointing to the user's avatar image.
keyword	String	Query String
friends_count	Integer	The number of users that this account is following (AKA their "followings").
followers_count	Integer	The number of followers that this account currently has.
description	String	A user-defined UTF-8 string describing their account.

Source: Twitter, 2014.

3.3 Data Pre-processing

Tweets usually include text with special symbols, such as the user mentions (@), hashtags (#), URLs (http links), and so on. Besides, because of the differences in writing systems between Thai and English, existing text pre-processing techniques could not be directly applied to this study's Thai sentiment classification system since there are no spaces between words in the Thai language. Moreover, a Thai Tweet has slang words, transliterated words, and emoticons, so data pre-processing was necessary before classification of the opinion. This process was subdivided into three steps as follows.

3.3.1 Text Extraction from Tweet

In this step, the Twitter-text java library developed by Goncalves (2014) was applied to extract the entities of Tweets from their text messages (URLs (http links), user mentions (@) and hashtags (#)). All the URLs (http links) were removed. For user mentions, the @ sign was removed along with the user mention except for when it matched one of the keywords. Lastly, the # sign was eliminated but the rest of the text after it was kept.

Example:

Input Tweet: "รู้สึกว่าจะเจอจุดอับสัญญาณเน็ต @TrueMoveH ตรงท่าเรือด่วนคลองแสนแสบ http://t.co/2JHzG5sKKv" (I feel that there is no signal at Port Saensaeb. @TrueMoveH http://t.co/2JHzG5sKKv)

Output: "รู้สึกว่าจะเจอจุดอับสัญญาณเน็ต TrueMoveH ตรงท่าเรือด่วนคลองแสนแสบ" (I feel that there is no signal at Port Saensaeb. TrueMoveH)

3.3.2 Text Pre-processing

In the text pre-processing step, four types of words were defined as abbreviations, transliterated words, slang words, and misspelled words, the requisite steps being to gather and organize words into their types. A total of 1,500 collected Tweets on three brands: AIS, DTAC, and TRUEMOVEH were used as an input source. After this, the method was subdivided into three steps as follows:

1) Create a new list file for each type.

2) Read and examine the text in the Tweet. If a word from one of the four types is found, it is added to the appropriate list file and assigned the original word.

3) Continue until 1,500 Tweets have been processed.

The texts that passed through the previous step was automatically checked for words defined as abbreviations, transliterated words, slang words, and misspelled words, after which they were replaced by expansions or the original words. Table 3.3 contains a sample and the number of words of each type discovered in this step.

Emoticons are domain and language independent (Khan et al., 2014) and have become an important token for social media content since they can express the feelings of the writer in the form of icons (Wikipedia, 2015a). For each Tweet, emoticons were assigned with a token label, as shown in Table 3.4. After this, the final step was to remove all of the digits.

Table 3.3 Example list of words in text pre-processing step

Words in Tweet	Original Word	Type	No. of Words in Type
พนง., พนง	พนักงาน (employee)	Abbreviations	19
สมาร์ทโฟน, สมาร์ตโฟน	Smartphone	Transliterated words	279
ปังมาก	ดีมาก (very good)	Slang words	217
ใบเส็ด, ใบเส็จ	ใบเสร็จ (receipt)	Misspelled words	17

Table 3.4 Examples of positive and negative emoticon sets

Positive Emoticons			Negative Emoticons		
Emoticons	**Meaning**	**Token Label**	**Emoticons**	**Meaning**	**Token Label**
:-) , :) , :D , :o) , :]	Happy	ehappyw	:-(, :(, >:[, :<	Sad	esadw
(^v^) , (^u^) , (^o^) , ^-^	Happy	ehappye	T-T , T^T , '_' , =_=	Sad	esade
:-D, 8-D, XD, =3, B^D	Laugh	elaughw	:'-(, :'(	Cry	ecryw

3.3.3 Thai Word Segmentation

The LongLexTo library was developed by the National Electronics and Computer Technology Center (NECTEC), Thailand (NECTEC, 1994). This library was constructed as a dictionary-based approach using the longest matching technique. Input text is scanned from left to right, and then the longest match with a word in the dictionary is selected along with any other matching words to improve the accuracy of word segmentation.

For the Thai word segmentation process, the LongLexTo java library was modified with a total of 42,833 words: 42,221 words from the Lexitron data dictionary of NECTEC and 612 words from related words in the domains of telecommunication and sentiment (Wiktionary, 2015). Texts passed on from the text-preprocessing step were automatically split into word tokens. For any English words (often included in Thai Tweets), conversion to lowercase was carried out. Lastly, other symbols were removed.

The flexibility of being able to add new words to the dictionary (including English words in common use) helped to improve the accuracy of the segmentation immensely. Moreover, Tweets are short (140 characters), and so the results of segmentation are better than if applied to longer texts.

3.4 Classification and Evaluation

The main objective of module was to identify the polarity of Thai opinion Tweets. The module was subdivided into two processes: (1) classification, and (2) evaluation.

3.4.1 Classification

Our classification system included the following steps:

1) Opinion filtering

2) Opinion polarity identification

In this research, the sentiment was divided into positive, neutral, and negative. The classification procedure for sentiment analysis from the TOM framework (Khan et al., 2014) was modified so as to be able to apply it to Thai Tweets. However, retrieved Tweets from the Twitter Search API are often combined with customers' opinions and news. Subsequently, the final results of customer's sentiment may be contaminated by this. Therefore, an opinion filtering process was added to classify Tweets that were really opinions from customers before using the classification procedure step from the TOM framework.

Step 1: Opinion filtering

Supervised learning techniques were applied in this task. A classifier was created to categorize Tweets based on opinion and non-opinion by using the WEKA java library (WEKA, 2014a). At first, an opinion model was developed using 1,000 messages (500 opinion and 500 non-opinion) for the training set. In this process, emoticons were used to improve accuracy and then strings were converted to word vectors by setting parameter TF-IDF (term frequency-inverse document frequency), as shown in equations (3-1) to (3-3) and removing stop words. Next, to construct the classification model, multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM techniques were used to create the opinion filtering model.

TF Transform in WEKA library, an information measure of the frequency of a word in a document, is defined as

$$\mathrm{TF} = \log_2\left(1 + f_{ij}\right) \tag{3-1}$$

where f_{ij} is the frequency of word i in document (instance) j.

IDF Transform in the WEKA library is given by

$$\mathrm{IDF} = f_{ij} * \log_2\left(\frac{num\ of Docs}{num\ of\ Docs\ with\ word(i)}\right) \tag{3-2}$$

where f_{ij} is the frequency of word i in document (instance) j. It is a measure of how much information the word provides; that is to say, whether the term is common or rare across all documents.

Next, TF-IDF is the product of two statistics: term frequency (TF) and inverse document frequency (IDF), as shown in equation (3-3):

$$\text{TF-IDF} = \mathrm{TF} * \mathrm{IDF} \tag{3-3}$$

Step 2: Opinion polarity identification

In order to identify the polarity of an opinion, only the procedure classification sentiment in the TOM framework was used combined with three classifiers: the enhanced emoticon classifier, the improved polarity classifier, and the SentiWordNet classifier. However, because the progress of the SentiWordNet project for the Thai language is 53.48% (approval of 117,659 senses) (Asian WordNet Project, 2007), it was not used to enhance identification of the polarity in Thai opinion Tweets in this research. As a result, this processing step was subdivided into two sub-steps as follows:

1) Emoticon classifier

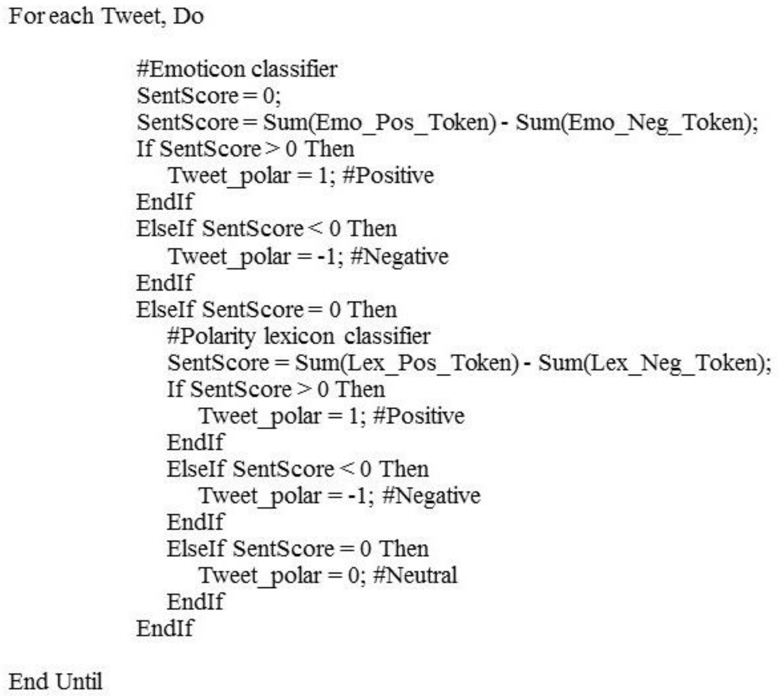
An emoticon is a short sequence of letters and symbols usually written to express a person's feelings or mood, and classification is based on sets of positive and negative emoticons. The emoticon was replaced by an emoticon token in the data-preprocessing module. A total of 140 emoticons from Wikipedia were used, 80 of which are tagged as positive and 60 are tagged as negative, with each emoticon token having the same weight (Wikipedia, 2015a). Positive and negative emoticon tokens in Tweets were counted and the sum recorded. Firstly, the sentiment score had an assigned value of 0. Each time a positive emoticon token was found, the score was incremented by one. On the other hand, if the emotion token was found in the negative set, the score was decreased by one. The sentiment of an opinion depended on the sum sentiment score: if the sum sentiment score was greater than zero, this constituted a positive opinion; if less than zero, this comprised a negative opinion; and if exactly zero, this signified a neutral opinion. The output was then passed to the polarity lexicon classifier step.

2) Polarity lexicon classifier

The polarity lexicon classifier uses a ‘bag of words’ approach where the set of positive and negative words were created from the Lexitron dictionary, Wiktionary, and Thai researcher (NECTEC, 1994; Wiktionary, 2015; Orathai Chinakarapong, 2014.). The total word count list was 506, which comprised 76 positive words and 430 negative words (sample sets of which are shown in Table 3.5) with each word having the same weight. Each word in a Tweet was checked for both positive and negative word sets to calculate the Tweet sentiment score. In the first step, the sentiment score has an assigned value of zero. Each time a positive word was found, the score was incremented by one. On the other hand, the discovery of a negative word meant that the score was decreased by one. At the end of the process, if the total sentiment score was greater than zero, then the opinion was marked as positive; if less than zero, the opinion was marked as negative; and a total score of zero classified the opinion as neutral. The algorithm for identifying the polarity of an opinion is shown in Figure 3.4.

Table 3.5 Positive and negative words examples

Positive Words		**Negative Words**	
ชอบ (like)	ดีมาก (very good)	แย่ (bad)	ห่วย (poor)
ประทับใจ (impressed)	ชมเชย (recommended)	แย่มาก (very bad)	กาก (the dregs)
รัก (love)	ปลื้ม (delighted)	เกลียด (hate)	เฮงซวย (inferior)



```
For each Tweet, Do

        #Emoticon classifier
        SentScore = 0;
        SentScore = Sum(Emo_Pos_Token) - Sum(Emo_Neg_Token);
        If SentScore > 0 Then
            Tweet_polar = 1; #Positive
        EndIf
        ElseIf SentScore < 0 Then
            Tweet_polar = -1; #Negative
        EndIf
        ElseIf SentScore = 0 Then
            #Polarity lexicon classifier
            SentScore = Sum(Lex_Pos_Token) - Sum(Lex_Neg_Token);
            If SentScore > 0 Then
                Tweet_polar = 1; #Positive
            EndIf
            ElseIf SentScore < 0 Then
                Tweet_polar = -1; #Negative
            EndIf
            ElseIf SentScore = 0 Then
                Tweet_polar = 0; #Neutral
            EndIf
        EndIf

End Until
```

Figure 3.4 The polarity identification algorithm of the ASTS

3.4.2 Evaluation

A confusion matrix, precision, recall, the F-measure, and accuracy were used as measures to evaluate the performance of the proposed method.

1) Confusion matrices

A confusion matrix contains information on the actual and predicted classes carried out by a classification system. It is a specific table layout that allows visualization of the performance of an algorithm. Table 3.6 shows the confusion matrix with each column representing the instances in the predicted class while each row represents the instances in the actual class.

Table 3.6 The confusion matrix

Dataset		**Predicted Class**		**Total**
		Class A	**Class B**	
Actual Class	Class A	tpA	eAB	tacA
	Class B	eBA	tpB	tacB
	Total	tpcA	tpcB	N

The entries in the confusion matrix have the following meaning in the context of this study:

(1) tpA and tpB are the numbers of correct classifications and are in the diagonal elements of the confusion matrix.

(2) tacA and tacB are the total number of actual instances of class A and class B, respectively.

(3) tpcA and tpcB are the total number of instances predicted as class A and class B, respectively.

(4) eBA and eAB are the numbers of incorrect classifications.

(5) N is the total number of instances.

2) Precision

Precision is defined as the fraction of true positives against all positive results (both true positives and false positives). The equation for calculating the precision of Class A is defined as

$$\text{Precision of A} = \frac{\text{tpA}}{\text{tpA+eBA}} \tag{3-4}$$

where tpA and eBA are the numbers of true and false positives for class A, respectively.

3) Recall

Recall is defined by the fraction of true positives against all actual classified positives (true positives + false negatives). The equation for calculating recall of Class A is defined as

$$\text{Recall of A} = \frac{\text{tpA}}{\text{tpA+eAB}} \tag{3-5}$$

where tpA and eAB are the numbers of true positives and false negatives for class A, respectively.

4) F-measure

The F-measure is defined as the harmonic mean between precision and recall, as shown in equation (3-6).

$$\text{F-measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{3-6}$$

5) Accuracy

The equation for calculating accuracy is defined by the proportion of true results (both true positives and true negatives) from all of the given data, as shown in equation (3-7).

$$\text{Accuracy} = \frac{\text{tpA+tpB}}{\text{N}} \tag{3-7}$$

3.5 Experimental Results and Evaluation

3.5.1 Data Collecting

To gather the Tweets (related to the three Thai mobile network operators mentioned earlier) submitted to the Twitter service from October 1, 2014 to March 31, 2015, data were collected using the Twitter search API which was configured to extract only Thai language Tweets by setting the parameter “lang=th” and excluding reTweets. Words related to the three mobile network operators were used as keywords for searching, as shown in Table 3.5. There were 72,661 Tweets captured in total and stored in a Tweets databank. 1,000 random Tweets from 1 October 2014 to 30 December 2014 were used to train the model, and 1,500 random Tweets from 1 January 2015 to 31 March 2015 were used to test the model. Tweets that contained more than one brand were excluded because the message may have included a comparative sentence, which is out of scope for this study.

Experiments were conducted on three different test datasets of random Tweets on the three companies being considered for analysis. Table 3.7 shows some examples of positive, negative, neutral, and non-opinion Tweets. The number of Tweets for each dataset are shown in Table 3.8.

Table 3.7 Examples of Tweets

Sentiment	**Keyword**	**Tweet**
Positive	AIS	ลอง net ใหม่ ais เร็วขึ้นนิดนึงยิ้มอ่อน
Negative	DTAC	threeg dtac คือ กากสุดเปิดแล้ว no service
Neutral	TRUEMOVEH	truemoveh ปรับ package fourg ใหม่ให้ มี fup เหมือนเดิมแล้วแฮะ
Non-opinion	DTAC	พนักงาน dtac ต้อนรับซีอีโอคนใหม่อย่างอบอุ่น iphonedroid

Table 3.8 Sample datasets

Dataset	Company Brand	No. of Tweets for Analyzing Sentiment Classification
Dataset 1	AIS	500
Dataset 2	DTAC	500
Dataset 3	TRUEMOVEH	500

The overall dataset for each class is given in Figure 3.5. Dataset 1, shown in Figure 3.5a, contained a total of 500 Tweets classified as 63 positive, 226 negative, 111 neutral, and 100 non-opinion. Figure 3.5b shows the distribution of dataset 2 which comprised a total of 500 Tweets classified as 56 positive, 248 negative, 96 neutral, and 100 non-opinion. Dataset 3 is shown in Figure 3.5c and consisted of a total of 500 Tweets classified as 72 positive, 256 negative, 72 neutral, and 100 non-opinion. Figure 3.5d shows the distribution of the overall dataset with a total of 1,500 Tweets classified as 191 positive, 730 negative, 279 neutral, and 300 non-opinion.

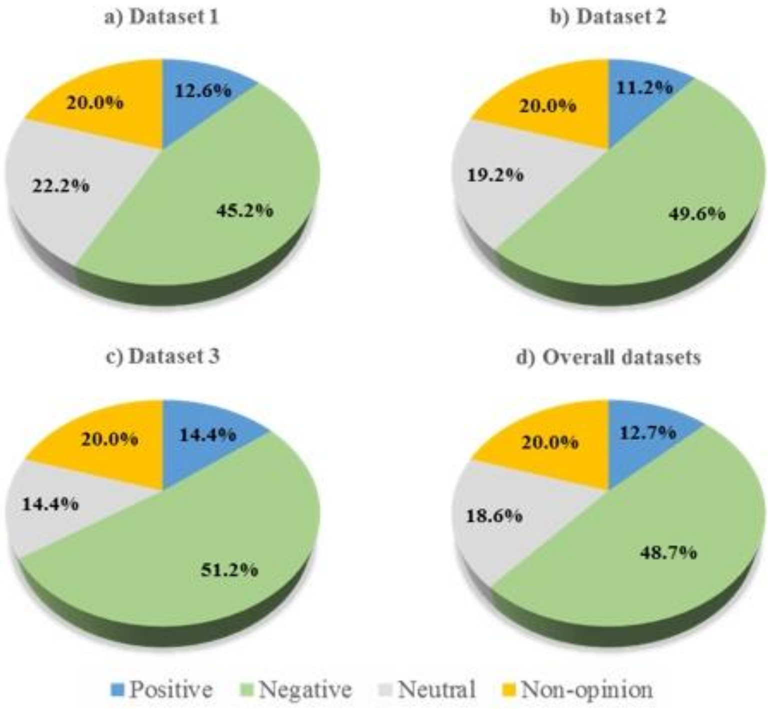


Figure 3.5 Distribution of the three separate and the overall datasets with non-opinion, positive, negative, and neutral sentiments

3.5.2 Results of the First Experiment

The experimental results were evaluated using a confusion matrix for precision, recall, F-measure, and accuracy. Supervised learning techniques were applied to perform opinion filtering classification. MNB and SVM techniques running under the WEKA java library environment were used to learn the classification model and evaluate the performance of classification. Besides, previous researches related to Thai opinion did not use emoticons in their analyses (Choochart Haruechaiyasak et al., 2010; Alisa Kongthon et al., 2013; Warunya Wunnasri, Thanaruk Theeramunkong and Choochart Haruechaiyasak, 2013). Nevertheless, the advantage of using emoticons is demonstrated in this research. The same training and testing datasets were used with the MNB and SVM techniques and trained by using two binary classification models for each technique with two classes: opinion and non-opinion. In the first model, all symbol characters were removed, and in the other, tokens were used to replace the emoticons and any other symbol characters were removed. Following this, string-to-word vectors were converted by setting parameter TF-IDF and removing stop words. For the training Tweets set, 500 opinion and 500 non-opinion random Tweets were prepared from the Tweet databank by manual classification. Moreover, for testing the performance of the process, 1000 random sample Tweets were taken and categorized as 500 opinion and 500 non-opinion.

The test results were based on 10-fold cross validation and shown in Table 3.9. The overall results of the techniques testing show that the MNB technique performed better than the SVM technique, as shown in Figure 3.6. Besides, the experimental results show that adding emoticon tokens (model 2) into the feature term improved the accuracy of opinion classification with both techniques. For the MNB technique, the accuracy improvement was 2.50%, and it was able to classify opinion and non-opinion with 91.10% accuracy. For the SVM technique, the accuracy improvement was 1.60 % and the overall accuracy was 86.60%.

Table 3.9 Results of opinion and non-opinion classification with the MNB and SVM techniques

Test Set		Total	MNB Technique			SVM Technique		
			Confusion Matrix		Accuracy	Confusion Matrix		Accuracy
			Opinion	Non-opinion		Opinion	Non-opinion	
Model 1 (no emoticons)	Non-opinion	500	463	37	88.60%	420	80	85.00%
	Opinion	500	77	423		70	430	
	Total	1,000	540	460		490	510	
Model 2 (emoticons)	Non-opinion	500	459	41	91.10%	408	92	86.60%
	Opinion	500	48	452		42	458	
	Total	1,000	507	493		450	550	

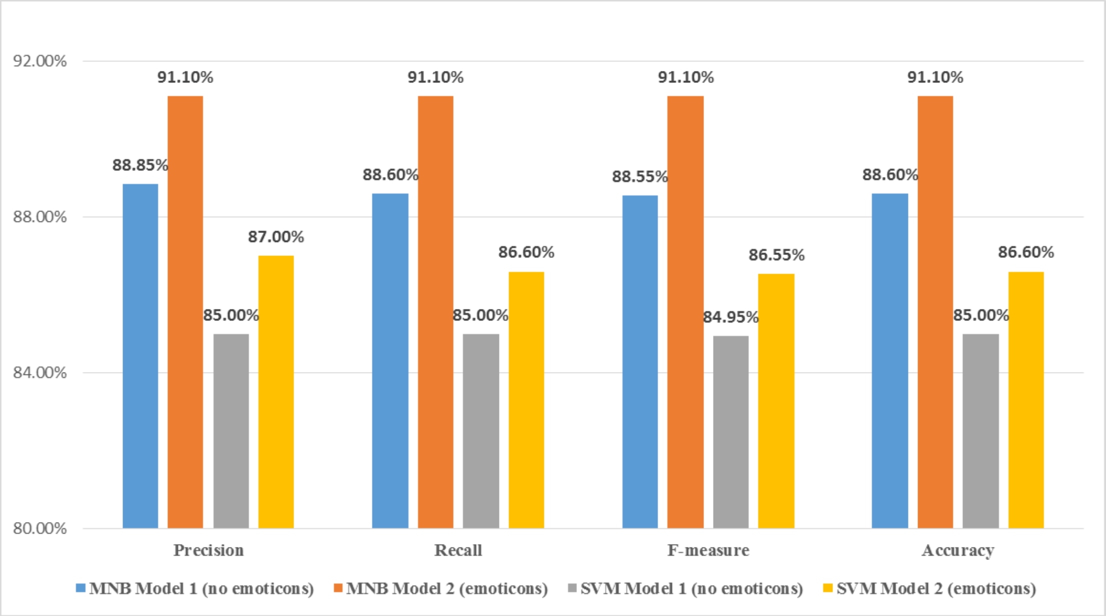


Figure 3.6 Precision, recall, and F-measurements for the MNB and SVM techniques

3.5.3 Results of the Second Experiment

The second experiment was to classify the sentiment of the Tweets. In the first experiment, the MNB technique showed better accuracy for opinion classification than the SVM technique, and so it was selected for use in the process of opinion filtering, and then to identify opinion polarity. Figure 3.7 shows the process extended from the TOM method (Figure 3.7a). The difference is in the classification module in that an opinion filtering module using the MNB algorithm was included (see Figure 3.7b).

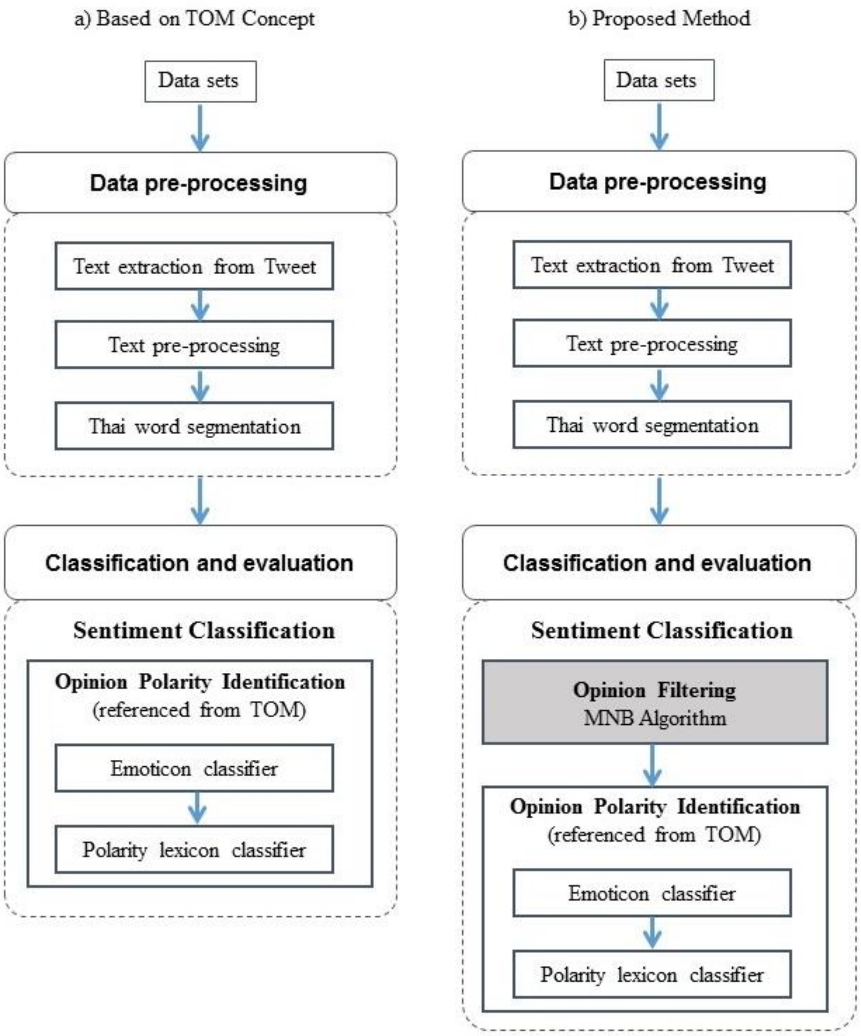


Figure 3.7 Processes of sentiment classification by TOM and the proposed method

The classification was run over the test datasets and processed each Tweet (the number of Tweets in each dataset are shown in Table 3.8). Each Tweet in a dataset was classified as either positive, negative, neutral, or non-opinion, for which the data distributions are shown in Figure 3.5. The same training and testing datasets were applied to evaluate the proposed method and TOM.

Tables 3.10 to 3.12 show the results for all three datasets. They demonstrate that the proposed method showed better performance for classification. The results of average precision, recall, F-measure, and accuracy of the proposed method for each dataset is shown in Figure 3.8. From the results, the proposed method was very effective in most cases with 84.80% accuracy (an improvement of 18.67%), 82.42% precision, 83.88% recall, and 82.97% F-measure.

Table 3.10 Results of sentiment classification for Dataset 1

Dataset 1		**Confusion Matrix**				**Total**	**Results**			
		Pos.	**Neg.**	**Neu.**	**Non.**		**Precision**	**Recall**	**F-measure**	**Accuracy**
TOM Concept	Pos.	47	4	12	N/A	63	51.6%	74.6%	61.0%	65.0%
	Neg.	7	192	27	N/A	226	88.5%	85.0%	86.7%	
	Neu.	12	13	86	N/A	111	44.8%	77.5%	56.8%	
	Non.	25	8	67	N/A	100	N/A	N/A	N/A	
	Total	91	217	192	N/A	500				
Proposed Method	Pos.	47	4	12	0	63	70.1%	74.6%	72.3%	84.2%
	Neg.	7	192	27	0	226	91.9%	85.0%	88.3%	
	Neu.	12	13	86	0	111	67.2%	77.5%	72.0%	
	Non.	1	0	3	96	100	100.0%	96.0%	98.0%	
	Total	67	209	128	96	500				

Note: Pos.: Positive, Neg: Negative, Neu.: Neutral and Non.: Non-opinion

Table 3.11 Results of sentiment classification for Dataset 2

Dataset 2		**Confusion Matrix**				**Total**	**Results**			
		Pos.	**Neg.**	**Neu.**	**Non.**		**Precision**	**Recall**	**F-measure**	**Accuracy**
TOM Concept	Pos.	47	3	6	N/A	56	54.7%	83.9%	66.2%	65.2%
	Neg.	6	213	29	N/A	248	88.8%	85.9%	87.3%	
	Neu.	9	21	66	N/A	96	37.9%	68.8%	48.9%	
	Non.	24	3	73	N/A	100	N/A	N/A	N/A	
	Total	86	240	174	N/A	500				
Proposed Method	Pos.	47	3	6	0	56	74.6%	83.9%	79.7%	84.0%
	Neg.	6	213	29	0	248	89.9%	85.9%	87.8%	
	Neu.	9	21	66	0	96	62.3%	68.8%	65.3%	
	Non.	1	0	5	94	100	100.0%	94.0%	96.9%	
	Total	63	237	106	94	500				

Table 3.12 Results of sentiment classification for Dataset 3

Dataset 3		Confusion Matrix				Total	Results			
		Pos.	Neg.	Neu.	Non.		Precision	Recall	F-measure	Accuracy
TOM Concept	Pos.	63	1	8	N/A	72	58.3%	87.5%	70.0%	68.2%
	Neg.	6	225	25	N/A	256	91.8%	87.9%	89.8%	
	Neu.	6	13	53	N/A	72	36.1%	73.6%	48.4%	
	Non.	33	6	61	N/A	100	N/A	N/A	N/A	
	Total	108	245	147	N/A	500				
Proposed Method	Pos.	63	1	8	0	72	78.8%	87.5%	82.9%	86.2%
	Neg.	6	224	25	1	256	93.7%	87.5%	90.5%	
	Neu.	6	13	53	0	72	59.6%	73.6%	65.8%	
	Non.	5	1	3	91	100	98.9%	91.0%	94.8%	
	Total	80	239	89	92	500				

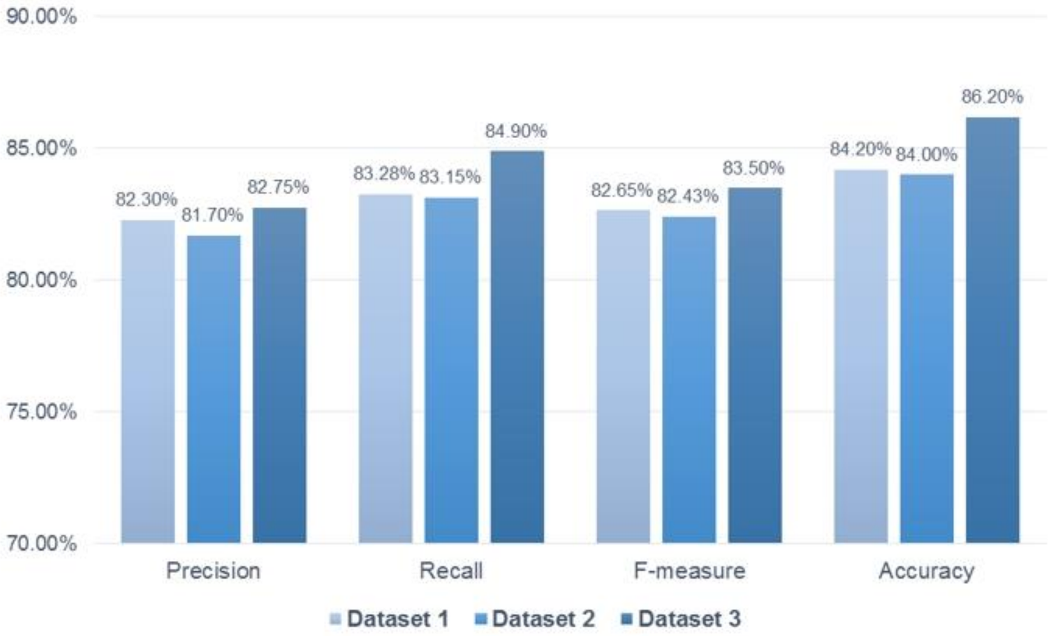


Figure 3.8 Average of precision, recall, F-measure and accuracy of the proposed method

From an in-depth analysis, non-opinion Tweets (errors) were found among the positive, negative and neutral categories, which made for low accuracy using the original TOM method (66.13%). An example of a non-opinion Tweet is “AIS จับมือ

CIMB เปิดตัวบริการใหม่ beat banking เพิ่มความสะดวก ในการทำธุรกรรมการเงินผ่าน MPAY" (AIS collaborates with CIMB bank to launch the beat banking service for creating convenient financial transactions through MPAY).

The results confirm that the proposed method significantly improved the original TOM-based method, which stems from this study's main contribution of adding the opinion filtering module. The graphical representation of improvements in the results between the two methods is illustrated in Figure 3.9, which clearly shows that opinion filtering helped to analyze Tweets more accurately. In addition, it was possible to make use of the filtering results, which are roughly classified into two groups of opinion and non-opinion Tweets, for other purposes or applications; for example, Tweets that are relevant to a particular company could be applied to brand monitoring and other business indicators.

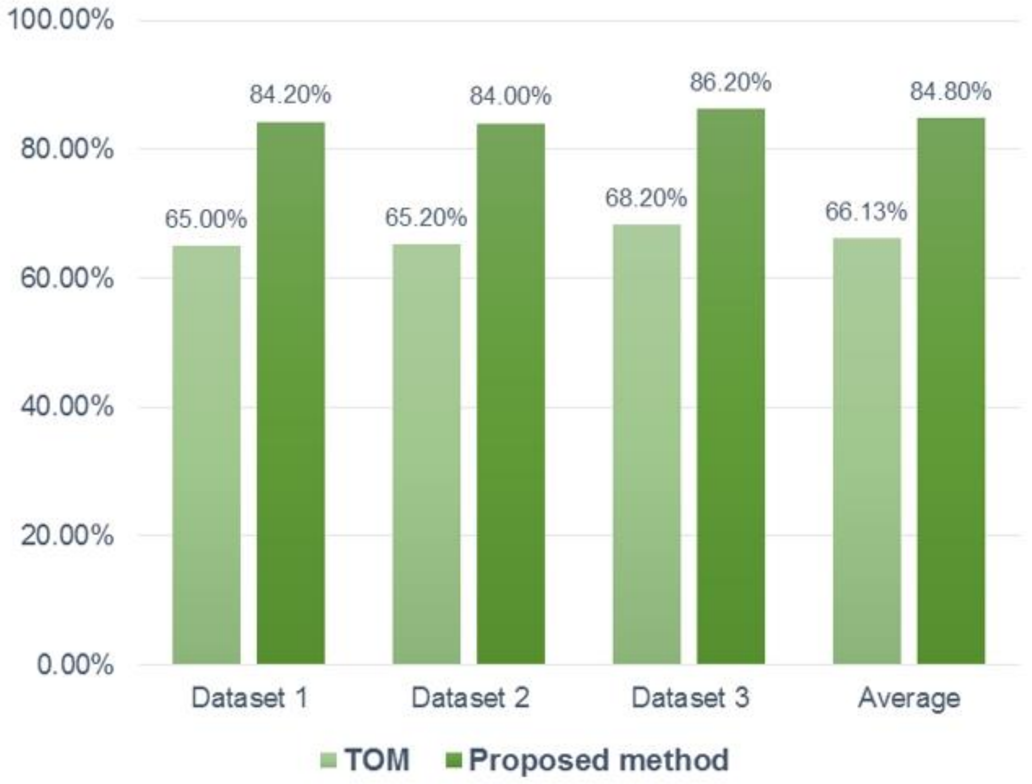


Figure 3.9 Performance of the proposed method for sentiment classification and its improvement over the TOM method

3.5.4 Error Analysis

To analyze errors, those test instances which had been misclassified were examined. The three major causes of errors can be summarized as word sense ambiguity, new slang words, and sarcasm. The first problem appears when a word contains many meanings depending on the context. For example, the word “เร็ว” (quick) when used with “สัญญาณ” (signal) will result in positive polarity, but on the other hand, when used with “ทวงเงิน” (debt collecting) will obtain negative polarity. To solve this problem, associated words were considered to help identify the polarity of the opinions.

The second problem is the generation of new slang words, which has become a new trend with Thais when using social media. This problem is made even worse in Thai because there are no spaces between words. For example, the word “แรงงง” (strong) can be split into “แรง งง” (strong confuse). Moreover, Thais use the original words with a new meaning. For example, the word “หอย”, which originally means “shellfish”, means lower speed in this context. The solution to this problem was to frequently add new words to and update existing ones in both the database of Thai word segmentation and the polarity lexicon. In addition, the context with respect to the particular business domain of interest was also taken into account.

The third problem is sarcasm, which makes it difficult to detect the polarity of opinions. It always results in the composition of a sentence with two possibly different polarities. For example Tweet “AIS 3G ครอบคลุมทุกจังหวัด แต่ที่บ้านเรา ไม่มีสัญญาณสักขีด” (AIS 3G cover all provinces but there is no signal at my home.) was considered to be a sarcastic sentence. These Tweets were mostly classified with a neutral opinion. However, it is still a difficult problem to solve and a challenging task in sentiment analysis (Choochart Haruechaiyasak et al., 2013).

3.6 Process Implementation

As mentioned in Chapter 2, 90% of Tweets included irrelevant mentions. Therefore, when applied all collected Tweets with ASTS, a data selection step is needed for selecting Tweets from commercial accounts and marketing agencies before passing them to the opinion filtering step. Figure 3.10 illustrates the process of using ASTS to create CI for supporting decision making. As the figure shows, the process can be divided into six steps as follows:

1) Data collection

2) Data pre-processing

3) Data selection

4) Opinion filtering

5) Opinion polarity identification

6) Results analysis

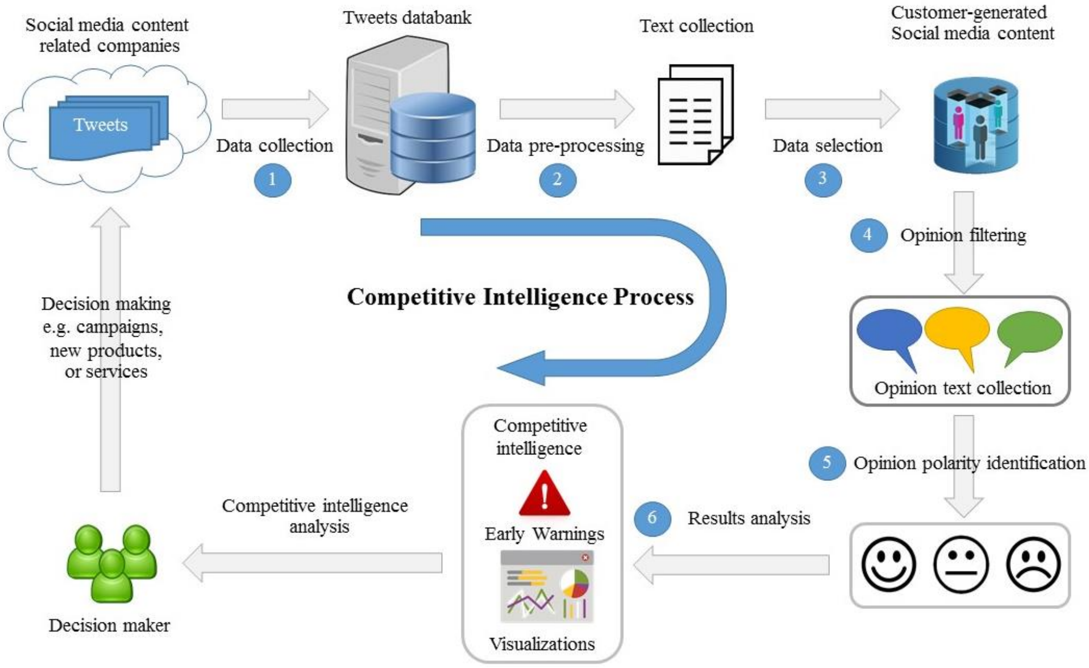

Social media content related companies
Tweets
Tweets databank
Data collection
1
Data pre-processing
2
Text collection
Data selection
3
Customer-generated Social media content
4
Opinion filtering
Competitive Intelligence Process
Decision making e.g. campaigns, new products, or services
Opinion text collection
Competitive intelligence
5
Opinion polarity identification
Early Warnings
6
Results analysis
Competitive intelligence analysis
Visualizations
Decision maker

Figure 3.10 Process of using ASTS for creating and using CI

The (3) data selection step is an additional step that is created for eliminating Tweets from commercial accounts and marketing agencies. The method was subdivided into two steps as follows:

1) To remove commercial Tweets, two fields of information in the Tweet table: screen_name and source were used in this step. The screen_name field was used to check whether a Tweet was from one of the official accounts of the three companies. If the official accounts from one of the three companies is found, Tweet is removed. The official accounts of the three largest mobile operator companies are as follows: ais_thailand for AIS, dtac for DTAC, and truemoveh for TRUEMOVEH.

2) For eliminating Tweets from marketing agencies, the source field of a Tweet (URLs) was compare with the name of the marketing agencies Twitterfeed, Hootsuit, and IFTTT. If the source field from one of the three marketing agencies is found, Tweet is removed.

The step (6) results analysis, it is divided in two parts of analysis as follow:

1) Descriptive analysis are used to describe the basic summaries of the results in the study. Some measures that are commonly used to describe a data set are measures of central tendency such as the mean, median, and mode, while measures of variability include the standard deviation (or variance), the minimum and maximum values of the variables.

2) SWOT analysis is used to evaluate company's strengths, weaknesses, marketing opportunities and potential threats to provide competitive insight into the potential and critical issues that impact the overall success of the business. In this research, the highly popular keywords from customers' Tweet are chosen for benchmarking the overall sentiment score. According to Bakshi, R. Kaur, N. Kaur and G. Kaur (2016: 454), the simple overall sentiment score that was calculated as

$$\text{Overall sentiment score} = \frac{\text{N(positive)}-\text{N(negative)}}{\text{N(positive)}+\text{N(negative)}} \qquad (3\text{-}8)$$

3.7 Summary

In this chapter, a method for analyzing Twitter sentiment by using both supervised learning techniques and lexicon-based techniques called ASTS was proposed. Experiments were conducted on social media data (Tweets) in the mobile network operator domain using the Twitter search API focused on Tweets in Thai. The results of testing the proposed method showed significant improvement on the basic concept of using the TOM framework. The average accuracy of ASTS was 84.80%, which shows great improvement (18.67%) over the original TOM framework (66.13%). In particular, this clearly shows that opinion filtering helped to analyze Tweets more accurately.

Moreover, the process of using ASTS for creating and using CI is also proposed. The results can be made use of in other applications (for example, Tweets that are relevant to a particular company), which could be useful for various applications such as brand monitoring, campaign monitoring, competitive analysis, and customer engagement. In Chapter 4, it provides results from using the proposed method for a case study of mobile network operators in Thailand.

CHAPTER 4

A CASE STUDY ANALYSIS

As an initial effort to implement the proposed method, ASTS was applied for analyzing the Twitter messages associated with three competing largest mobile operators in Thailand (AIS, DTAC, and TRUEMOVEH) and creating CI. Tweets were collected using the Twitter search API which was configured to extract only Thai language Tweets by setting the parameter "lang=th" and excluding reTweets. Table 3.1 shows words associated with the three rivals used as keywords for searching. Tweet data include various fields of information (as shown in Table 3.2) that can help a researcher recognize critical clues and retrieve further data qualification. The collecting program ran every day to collect Tweets, 72,661 in total over a period of six months from October 1, 2014 to March 31, 2015. Figure 4.1 displays a bar chart and a linear trend line in a series of retrieve Tweets. This trend line shows that the use of Twitter in Thailand was increasing between these dates.

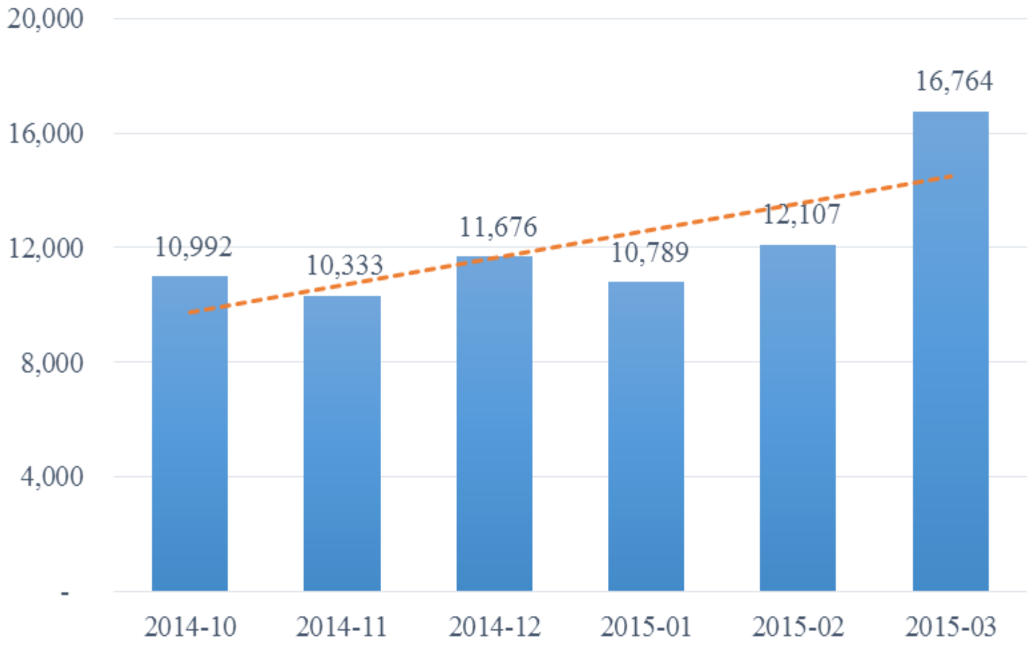


Figure 4.1 Tweets collecting over the six month data collection period

4.1 Results Analysis

After collecting 72,661 Tweets and then pre-processing them, those that did not mentioned the name of one of the three mobile operator companies were removed, leaving 46,909 Tweets. The next step, the data selection is needed for filtering Tweets from commercial accounts and marketing agencies (examples are shown in Table 4.1). Then, the filtering opinion step from ASTS was used to classify Tweets that were real opinions from customers, as shown in Table 4.2.

Table 4.1 Examples of commercial Tweets

Mobile Operator Company	**Commercial Tweet**
AIS	AIS Tweet: เอไอเอส ลดค่าเครื่องอลังการ! iPhone เริ่มต้นเพียง 5,990 บาท เท่านั้น http://t.co/J6FPdatEjt
DTAC	[PR] Microsoft จับมือ Dtac มอบแพคเกจสำหรับ Nokia Lumia 730 http://t.co/LuDrFUNWif
TRUEMOVEH	ทรูมูฟ เอช มอบสิทธิพิเศษให้ลูกค้าแบบเติมเงิน และซิมโซเชียลไม่อั้น http://t.co/GL0QhTalm2

Table 4.2 Number of remaining Tweets

Mobile Operator Company	**Mention Brand Tweets**	**Removal by Official Accounts Check**	**Removal of Marketing Sources**	**Removal by Opinion Filter**	**Remaining Tweets**
AIS	21,009	120	2,396	7,295	11,198 (53.30%)
DTAC	16,073	518	1,664	5,456	8,435 (52.48%)
TRUEMOVEH	9,827	320	1,047	3,336	5,124 (52.14%)
Total	46,909	958	5,107	16,087	24,757 (52.78%)

As a result, about 47% of the total number of Tweets (excluding reTweets) were eliminated. The results show that most of Tweets were news and commercial Tweets. In summary, the total number of Tweets was reduced from 46,909 to 24,757, consisting of 11,198 AIS, 8,435 DTAC, and 5,124 TRUEMOVEH Tweets. Then, identify the polarity of Tweets by using the process opinion polarity identification from ASTS.

4.1.1 Opinion's Tweet Volume Analysis

The time series visualization graph in Figure 4.2 clearly shows differences in volume among the big three companies in the telecommunication industry. The Tweet volume related to AIS was generally higher than DTAC and TRUEMOVEH that was accordant with market shares. Table 4.3 shows that the mean weekly volume for AIS (*m* = 425.48) was larger than DTAC (*m* = 322.52) and TRUEMOVEH (*m*=196.20).

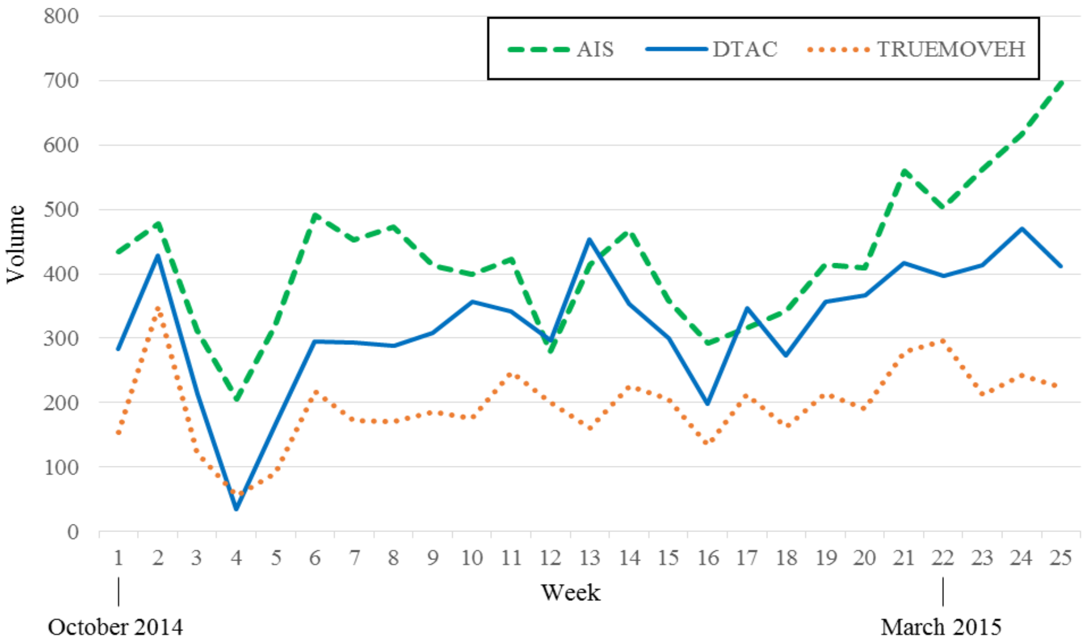


Figure 4.2 A comparison of Tweet opinion volumes

Table 4.3 Descriptive statistics of the Tweet volumes of the three companies

Mobile Operator Company	**Market Shares (Q1, 2015)**	**Weekly Volume Mean**	**N**	**SD**	**Min.**	**Max.**
AIS	46.52%	425.48	25	111.92	205	696
DTAC	28.50%	322.52	25	97.84	34	471
TRUEMOVEH	24.26%	196.20	25	62.64	56	350
Total		314.73	75	131.72	34	696

4.1.2 Sentiment Analysis

Because the scope of ASTS (86.20% accuracy) involves detecting the polarity of Thai Tweets based on one brand mention, Tweets that mentioned only one brand were selected. This process classified Tweets in three categories: positive, neutral, and negative, an example of which is shown in Figure 4.3.

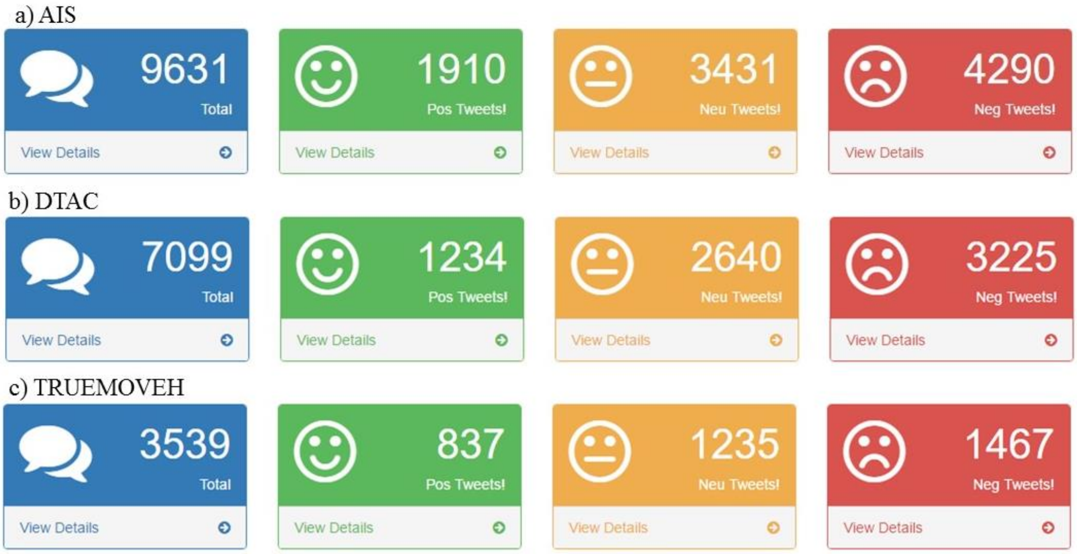


Figure 4.3 The summarize screen of total Tweets with classified categories

Table 4.4 shows the opinion summary of the three companies. There are 20,269 in total, 9,631 Tweets for AIS (47.52%), 7,099 for DTAC (35.02%), and 3,539 for TRUEMOVEH (17.46%).

Table 4.4 Summary of the opinion mining results

Company	Tweets (One Brand)	Positive	Neutral	Negative	Overall Sentiment
AIS	9,631/20,269	1,910/9,631	3,431/9,631	4,290/9,631	Negative
	(47.52%)	(19.83%)	(35.62%)	(44.54%)	-0.38
DTAC	7,099/20,269	1,234/7,099	2,640/7,099	3,225/7,099	Negative
	(35.02%)	(17.38%)	(37.19%)	(45.43%)	-0.45
TRUEMOVEH	3,539/20,269	837/3,539	1,235/3,539	1,467/3,539	Negative
	(17.46%)	(23.65%)	(34.90%)	(41.45%)	-0.27
Total	20,269	3,981	7,306	8,982	

Figure 4.4 shows a sentiment comparison for the three companies. For positive sentiment, TRUEMOVEH was the highest (23.65%), then AIS (19.83%), and finally DTAC (17.38%). For negative sentiment, DTAC was the highest (45.43%), followed by AIS (44.54%), and TRUEMOVEH (41.45%). The overall sentiment scores of the three companies were negative with the score for TRUEMOVEH (-0.27) being slightly better than AIS (-0.38) and DTAC (-0.45).

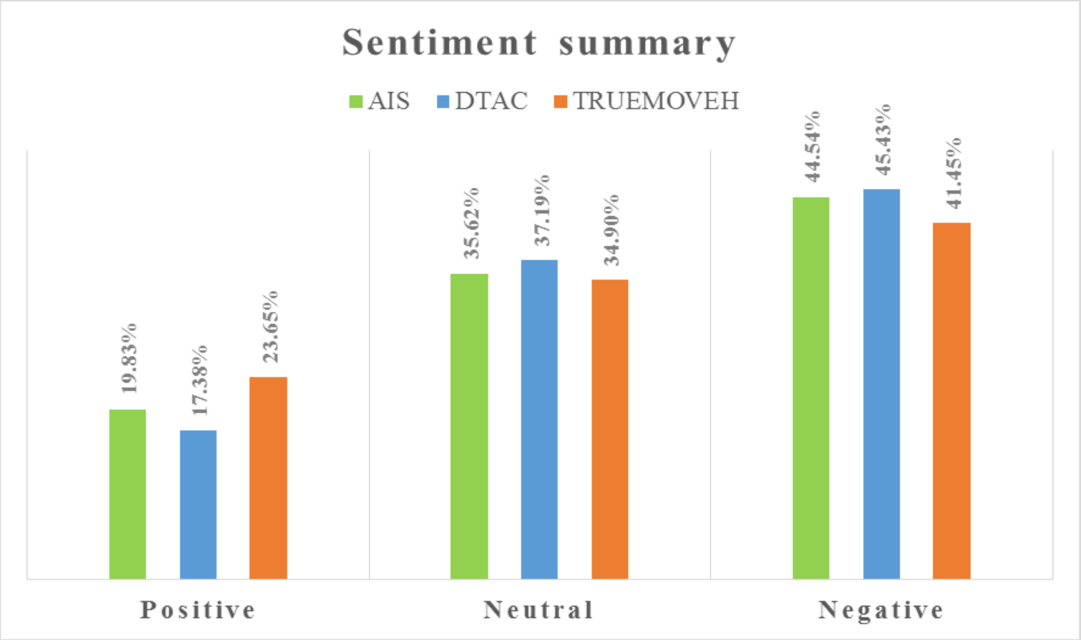


Figure 4.4 Summary of sentiment for the three companies

The results show that the customers of all three companies were not satisfied with their products and services, which could be viewed as a challenge for all three companies to quickly identify dissatisfaction and then correct the problems. In particular, DTAC (with a lower overall sentiment score) may have more problems than AIS and TRUEMOVEH.

As background information, Figure 4.5 illustrates the movement of stock prices for the three competing companies. The stock price movement of AIS and TRUEMOVEH were trending upward, while the stock price movement of DTAC was in a downward trend, which corresponds well to the sentiment calculated in this study and from overall VOC in social media.

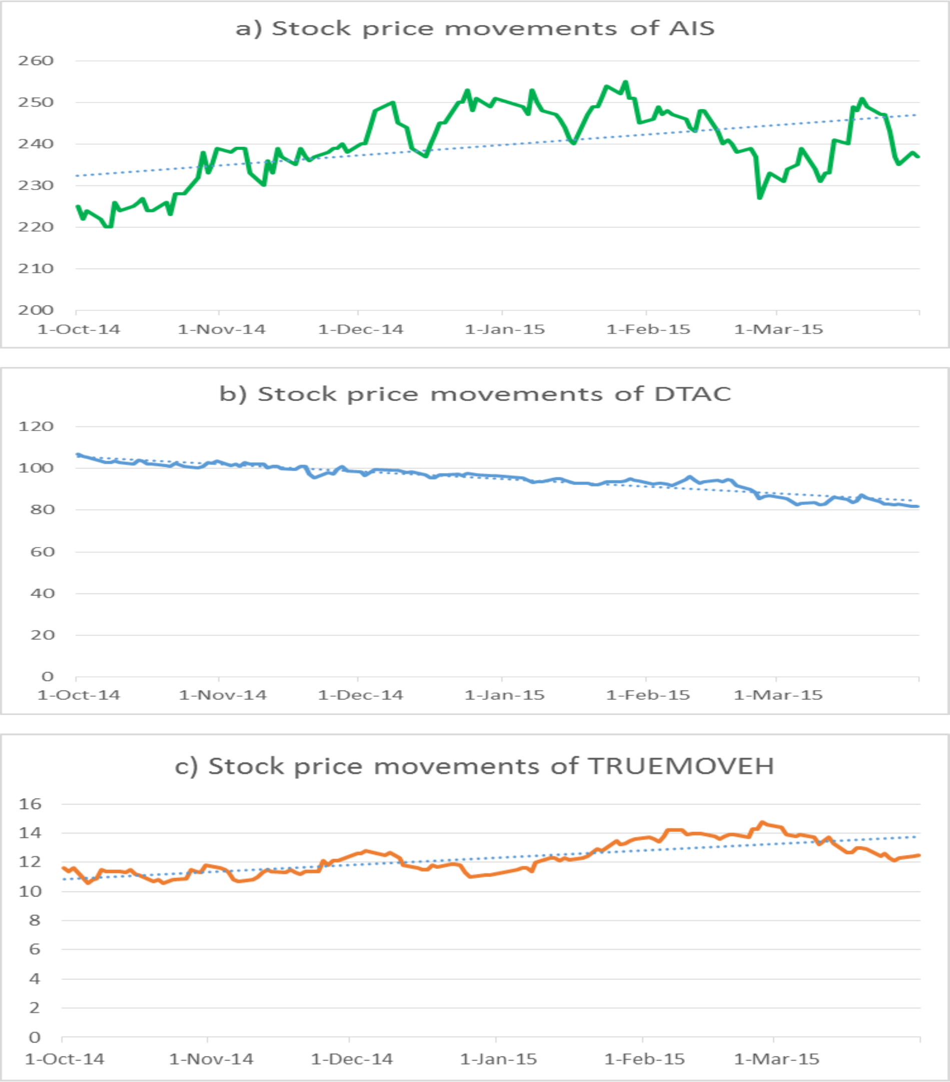


Figure 4.5 Stock price movements of the three companies

This situation may have been caused by the effect of widespread public sentiment expressed on social media (such as Twitter) that investors can monitor rapidly. Hence, this may be the reason for the continuous selling off of DTAC stock. However, this point is out of the scope of this study, so further research needs to be carried out to determine the relationship between customer sentiment on social media and stock price.

4.1.3 Measuring Brand and Understanding Conversation Sentiment

The overall opinion volume and sentiment scores for the big three mobile operators in Thailand over a period of six months are illustrated in Figures 4.6-4.8. The results show that their customers used Twitter to express their experience more often than in the past and that their sentiment is mostly negative. Figure 4.6 shows that the trend in sentiment score of AIS is down and was particularly bad in March 2015 (-0.49). Figure 4.7 shows that the sentiment score of DTAC in January 2015 was extremely bad (-0.60), then improved in February (-0.41) and March (-0.45). For TRUEMOVEH (Figure 4.8), the trend in sentiment score was downward overall and became exceptionally bad in March 2015 (-0.43).

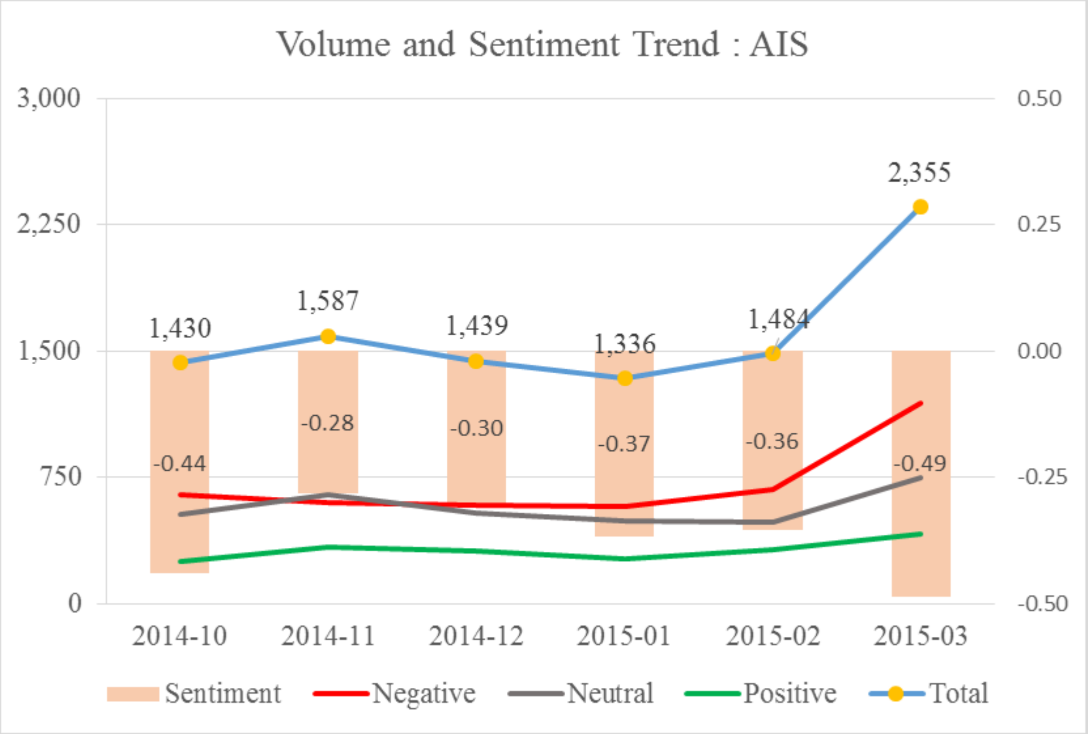


Figure 4.6 AIS volume and sentiment trend

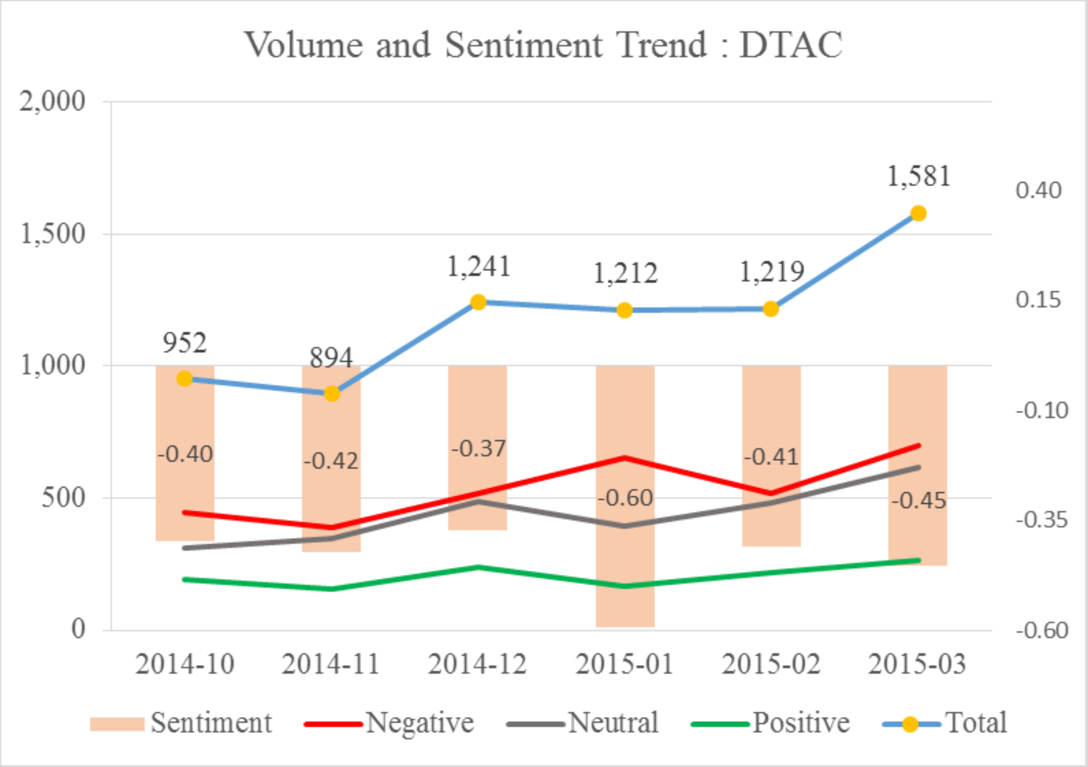


Figure 4.7 DTAC volume and sentiment trend

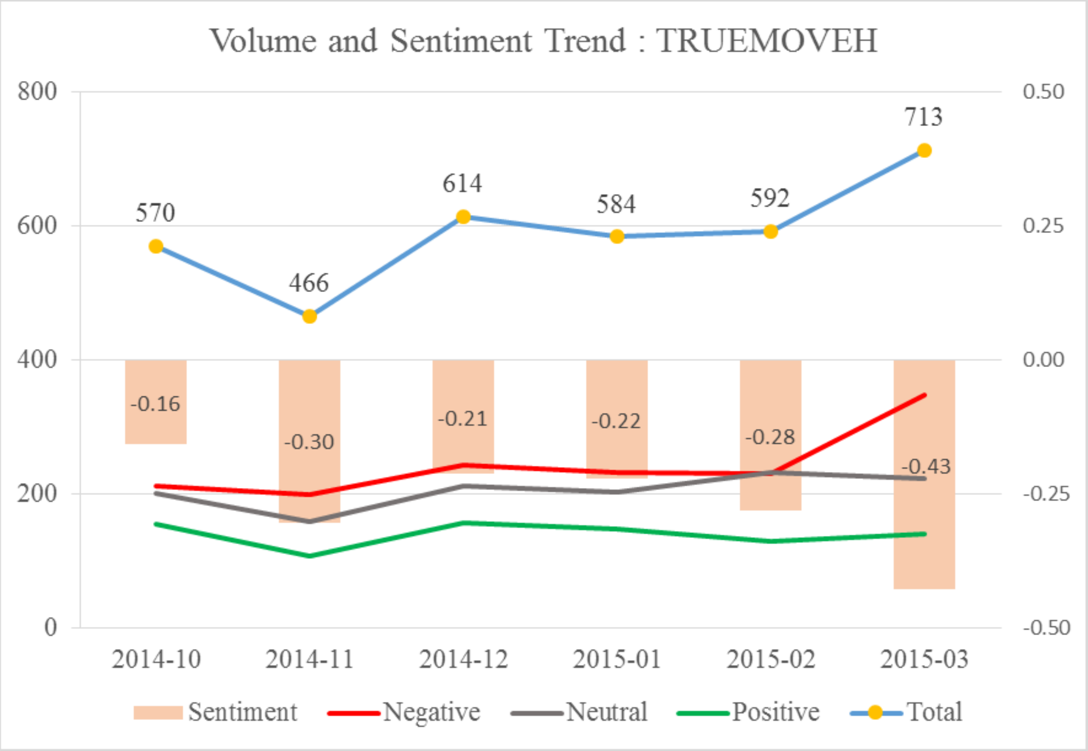


Figure 4.8 TRUEMOVEH volume and sentiment trend

It can be concluded that dissatisfaction of customers of all three companies were increasing, which indicates a problem. Hence, AIS, DTAC, and TRUEMOVEH should concentrate on VOC from social media in order to respond and solve the problem to reduce dissatisfaction. Moreover, they should be focus on popular keywords for buzz tracking and early alerting.

4.1.4 Discovery Trending Topics

In addition to the overall volume and sentiment trend analysis, the volume and sentiment trend analysis at the individual product or service level is of interest since all three companies are direct competitors and often offer the same type of products and services. In addition, the keywords that are always mentioned in the Tweets of each company were examined, as shown in Figures 4.9-4.11.

The method for selecting categories of this study is used the data-driven concept by choosing the top same noun keywords. Hence, four highly popular keywords from retrieved Tweets were chosen: สัญญาน (net), วายฟาย (wifi), โปรโมชั่น (promotion), and ย้ายค่าย (switching). However, the categories result was also accordant with the concept of market mix 4P's, product, price, place and promotion. In addition to the four keywords, keyword "พนักงาน (employee)" was considered as being very important for service businesses.

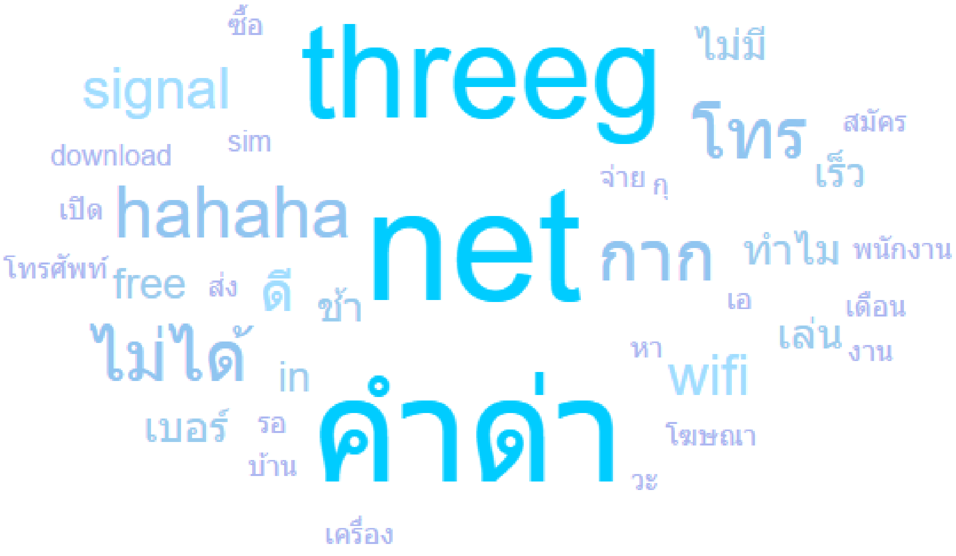


Figure 4.9 Word cloud from keyword frequencies for AIS

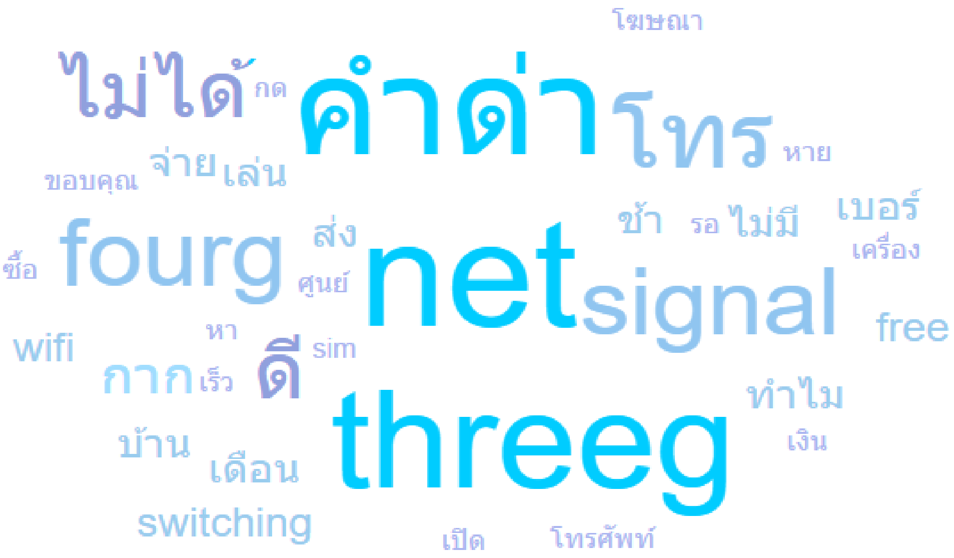


Figure 4.10 Word cloud from keyword frequencies for DTAC

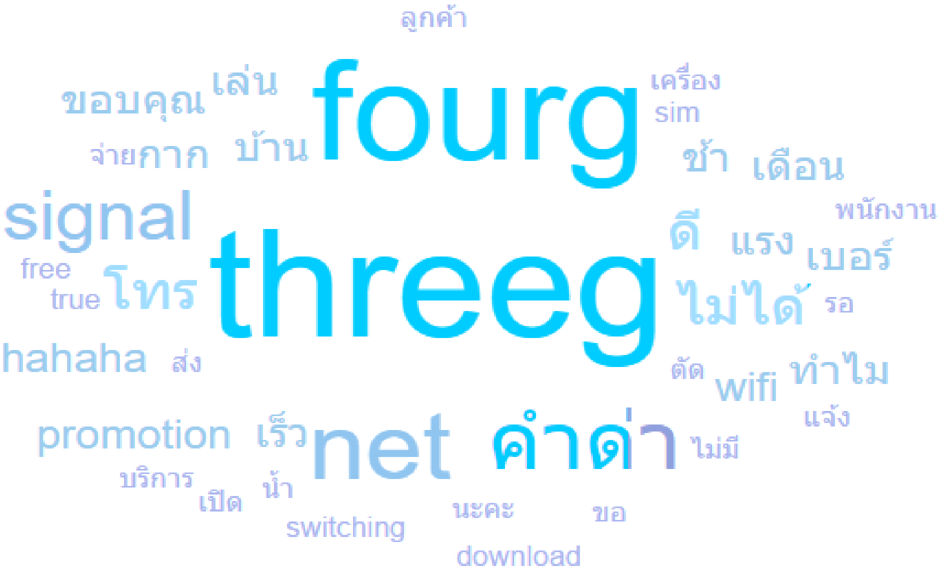


Figure 4.11 Word cloud from keyword frequencies for TRUEMOVEH

Tables 4.5-4.7 show the keyword mentions as well as the number of positive, negative, and neutral Tweets related to the keywords of interest for the three companies. The most of all keyword mentioned of three companies were in the negative sentiment. For AIS, overall sentiment about product ("net") was very bad, -0.71. Table 4.6 shows that the overall sentiment of DTAC about product ("net") was also bad, -0.68. For TRUEMOVEH, overall sentiment about employee was very bad, -0.75.

Table 4.5 AIS opinion Tweets

Mention Keyword	**No. of Mentions**	**Sentiment**			
		Positive	**Neutral**	**Negative**	**Overall**
net	2,061/9,631 (21.40%)	237 (11.5%)	432 (20.96%)	1,392 (67.54%)	Negative -0.71
wifi	589/9,631 (6.12%)	152 (25.81%)	231 (39.22%)	206 (34.97%)	Negative -0.15
promotion	446/9,631 (4.63%)	102 (22.87%)	123 (27.58%)	221 (49.55%)	Negative -0.37
switching	264/9,631 (2.74%)	40 (15.15%)	89 (33.71%)	135 (51.14%)	Negative -0.54
employee	177/9,631 (1.84%)	61 (34.46%)	59 (33.33%)	57 (32.2%)	Positive 0.03

Table 4.6 DTAC opinion Tweets

Mention Keyword	**No. of Mentions**	**Sentiment**			
		Positive	**Neutral**	**Negative**	**Overall**
net	1,225/7,099 (17.26%)	153 (12.49%)	268 (21.88%)	804 (65.63%)	Negative -0.68
wifi	217/7,099 (3.06%)	38 (17.51%)	65 (29.95%)	114 (52.53%)	Negative -0.50
promotion	350/7,099 (4.93%)	68 (19.43%)	145 (41.43%)	137 (39.14%)	Negative -0.34
switching	241/7,099 (3.39%)	38 (15.77%)	80 (33.2%)	123 (51.04%)	Negative -0.53
employee	116/7,099 (1.63%)	32 (27.59%)	36 (31.03%)	48 (41.38%)	Negative -0.20

Table 4.7 TRUEMOVEH opinion Tweets

Mention Keyword	No. of Mentions	Sentiment			
		Positive	Neutral	Negative	Overall
net	546/3,539 (15.43%)	107 (19.6%)	116 (21.25%)	323 (59.16%)	Negative -0.50
wifi	187/3,539 (5.28%)	34 (18.18%)	64 (34.22%)	89 (47.59%)	Negative -0.45
promotion	164/3,539 (4.63%)	31 (18.9%)	64 (39.02%)	69 (42.07%)	Negative -0.38
switching	103/3,539 (2.91%)	22 (21.36%)	35 (33.98%)	46 (44.66%)	Negative -0.35
employee	78/3,539 (2.20%)	7 (8.97%)	23 (29.49%)	48 (61.54%)	Negative -0.75

A radar chart is a useful way to display many observations with an arbitrary number of variables. It is suitable to use for comparing the interesting categories. Based on the top five mentioned keywords analysis, a radar chart was used to compare each keyword mentioned for each company. Figures 4.12 - 4.14 contain the radar charts of sentiment scores for the top five mention keywords.

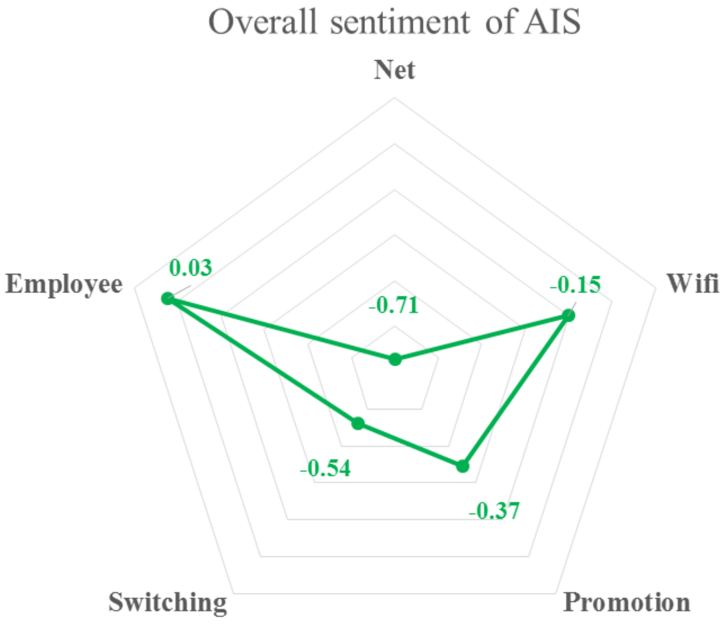


Figure 4.12 Radar chart of the five keyword mentions in Twitter data for AIS

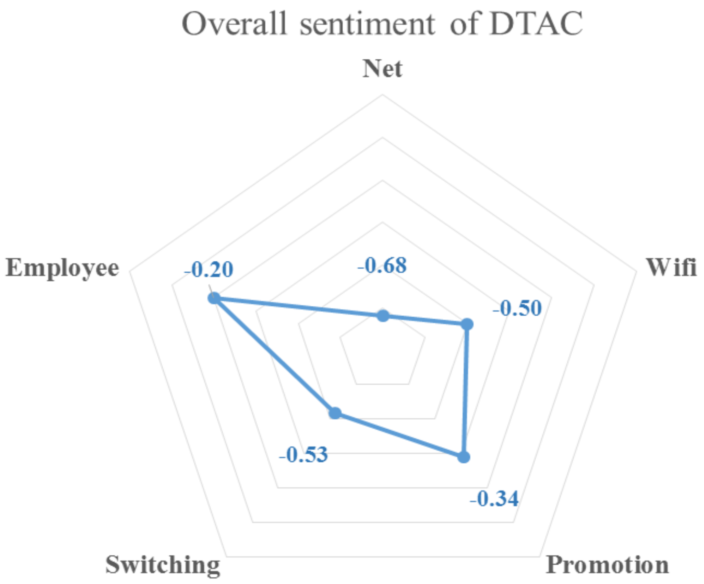


Figure 4.13 Radar chart of the five keyword mentions in Twitter data for DTAC

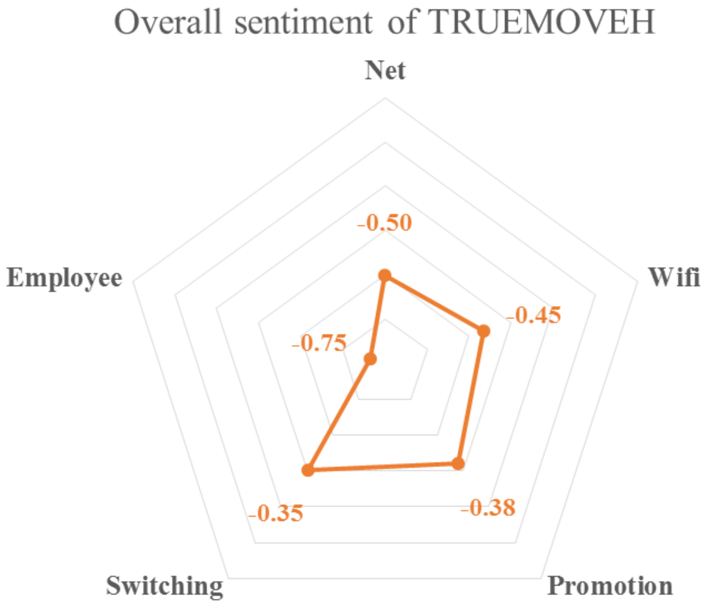


Figure 4.14 Radar chart of the five keyword mentions in Twitter data for TRUEMOVEH

Figure 4.12 shows the sentiment for AIS, and it was found that "employee (พนักงาน)" was obviously the best one since its overall sentiment score was positive (0.03). However, "Net (สัญญาน)" was flagged as a major concern since its overall sentiment was extremely low (-0.71). Figure 4.13 contains the radar chart for DTAC in which all of the sentiment scores for the five keyword mentions are in the negative zone. Although “employee (พนักงาน)” was the best one for DTAC, its sentiment score was still negative (-0.20). Moreover, it was found that “Net (สัญญาน)” was also flagged as a major problem since its sentiment was particularly low (-0.68).

For TRUEMOVEH (Figure 4.14), “switching (ย้ายค่าย)” had a higher sentiment score than the other keywords even though it was still negative (-0.35). The mention keyword “employee (พนักงาน)” was indicated as a significant problem for TRUEMOVEH since its overall sentiment was the lowest in all of the charts (-0.75), especially since AIS and DTAC showed much better sentiment for this keyword. Therefore, it is recommended that TRUEMOVEH should quickly resolve this problem by enhancing employee satisfaction.

In order to find out the cause of problems indicated by Twitter data, network graph plotting is used to illustrate the co-occurrence of word, as shown in Figure 4.15. An example case (“employee (พนักงาน)” for TRUEMOVEH) is demonstrated in which the method was subdivided into five steps as follows:

1) Select negative Tweets that mention “employee (พนักงาน)” for TRUEMOVEH (48 Tweets).

2) Calculate the frequency of words and select the top twenty words with the highest frequencies.

3) Create a document matrix file.

4) Calculate the co-occurrence of the twenty words.

5) Plot network graph of the words.

Figure 4.15 illustrates the twenty words that most often occurred with “employee (พนักงาน)”. The size of a circle represents the word frequency and the width of a line indicates the frequency of co-occurrence between one of the words and another. The diagram infers that the problem concerning employee may occur at the TRUEMOVEH shops. An example Tweet that has the keyword “employee (พนักงาน)” for TRUEMOVEH with negative sentiment is shown below:

Screen_name (Owner): NickiieZ092

Date: 15-01-25 13:41:26

Tweet content: @TrueMoveH พนักงานก็แย่ไปนะครับ หน้าไม่รับแขก ไม่มีการประสานงาน ตอบแค่ว่า ผมไม่รู้ คนละทีม เสียความรู้สึกมาก (@truemoveh employees are very bad, look displeased. Don't collaborate with any team to solve my problem, say only don't know. Very bad.)

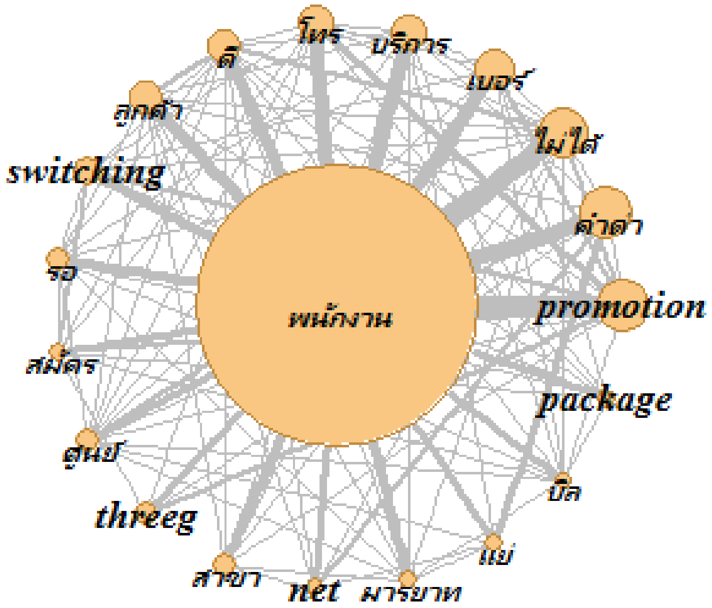


Figure 4.15 Co-occurrence words of “Employee (พนักงาน)” of TRUEMOVEH

This case illustrates that a customer (NickiieZ092) went to the shop and received poor service from its employees. It is obvious that this message was expressed with real feeling. To solve this complaint, TRUEMOVEH should contact NickiieZ092 quickly and resolve his/her problem, then ask him/her about his/her experience at that shop to discover the facts. Finally, this information should be used as input to improve the customer service process.

Another piece of information from a Tweet that is very useful is screen_name. It was found that there were 6,431 screen_name (username on Twitter) variations that mentioned AIS, 4,598 for DTAC, and 2,296 for TRUEMOVEH. Table 4.8 shows the top twenty screen_name variations that posted for each company.

Table 4.8 Top twenty posting users' screen_name of the three companies during the six month data collection period

No.	AIS		DTAC		TRUEMOVEH	
	Screen_name	**No.Post**	**Screen_name**	**No.Post**	**Screen_name**	**No.Post**
1	mckung	175	mckung	106	mckung	38
2	Nonny_Ramones	57	iPlugz	66	kafaak	34
3	tannce	32	theerayuthj	47	iBabe	33
4	minxkung	30	kafaak	44	Crysty_Choi	23
5	iPlugz	29	Sakeuth	35	BFood1717	22
6	Tamagotnc	27	Junno_botTH	34	noot010	19
7	Mnakin	24	iBabe	29	SnackPro	19
8	SnackPro	23	darkmasterxxx	28	FireDept33	16
9	fuckyeahohyeah	21	Nexttime2gether	23	Sakeuth	15
10	chocopair	20	_bqxz	21	imtaiki	15
11	FONGFODZS	19	waans_	20	iamafeave	15
12	dj_n4	17	imtaiki	17	JIA_Samrongthap	14
13	BehemortHz	17	MacStroke	17	Mashiiro	13
14	gururv33	17	suebsak1	16	waans_	13
15	Nechigawara	17	nomi_nee	16	jokermaster619	13
16	Sukannika_28	17	ubmint	15	nuttoday	12
17	kafaak	16	oathth	14	AppDisqus	12
18	Pimry_zaa	16	SnackPro	13	phisitcute	11
19	AdmOd	16	Nicky_sass	13	Koa_Ka	11
20	icez	15	iCabq	13	Saipaannana	11

Username "mckung" was at the top for posting on all three companies. He/she used the services from all of them, and always post Tweets about them with negative sentiment. Example Tweets from "mckung" are as follows:

Example 1: ais แถว บ้าน ที่ คำด่า เสา นิดเดียว signal ก็ หาย ไม่ได้ เป็น เพราะ mobilephone กาก แต่ ais กาก เอง เครียด เลย วะ (Signal pole AIS is too small and low signal or my mobile phone is bad, be nervous.)

Example 2: dtac net รั่ว จาก ระบบ ได้ ไง วะ เครื่อง ไม่ได้ เปิด ใช้ data เจอกัน หลาย คน เลย เอ๊ะ หรือ วิธี หากิน ใหม่ (DTAC I don’t open data mode but net leak. Many people found this too. Or this is a new business model?)

Example 3: แต่ truemoveh ลูกค้า น้อย สุด ตั้งแต่ ไหน แต่ ไร threeg ก็ กาก เหมือนเดิม เครียด สักครู่ (TRUEMOVEH has the smallest number of customers but threeg is too bad. Be serious.)

The result shows that these information from social media supplies information on customers such as their experiences when using products and services. In addition, companies could use screen_name to contact and engage with its customers. Hence, their marketing departments could send web survey links to ask customers about their level of satisfaction with their experiences and request more information such as mobile phone number to integrate offline and online data. This is the challenge for this era, for integration of all physical channels (offline) and digital channels (online) in a unified customer experience (such as the Omni-channel business model).

Moreover, this information also contains details on customers of competitors. Company can also use screen_name for connecting to customers of competing companies and offer better promotions or services. For example, if this system found a Tweet from an AIS customer that mentions switching with negative sentiment, DTAC or TRUEMOVEH might be able to attract the AIS customer by sending a Tweet with content on a promotion for switching to the screen_name. This would be a simple tactic for competing.

4.1.5 Competitive Intelligence

Benchmarking against competitors produces essential information for the CI decision-making process and is also a major component of CI. A comparison of Tweet volumes from the different mobile operator companies in the study: AIS, DTAC, and TRUEMOVEH is shown in Figure 4.16. By comparing the number of Tweet mentions in Figure 4.16, it was found that AIS was mentioned the most over the six month period.

Hence, a comparison of Tweet mentions can provide a clear understanding of customers’ attention on the different companies within the same industry.

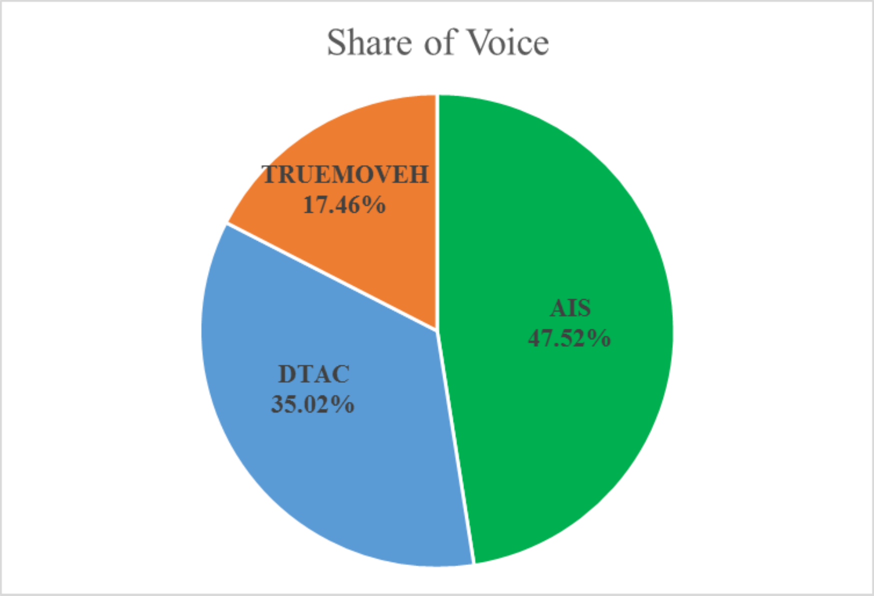


Figure 4.16 Comparison of the volume of Twitter mentions for the three companies

A perceptual map was used to represent the position of customers’ sentiment for the three companies during the entire period of data collection using Tweet volumes, Tweet sentiment, and Twitter followers. Figure 4.17 clearly illustrates that all three companies had a low level of customer satisfaction (negative sentiment). The customers mentioned AIS and DTAC in Tweets more often than TRUEMOVEH. The other dimension shows the number of followers of each company (denoted by circle width in Figure 4.17). It shows that the number of followers of DTAC was higher than the other two companies. This point is very interesting in that why does DTAC have more followers than AIS (which is the market share leader)? Therefore, this information could also be added to the input of the decision-making process. In addition, it can be concluded that the products and services of the three companies should make some improvements to increase customer satisfaction.

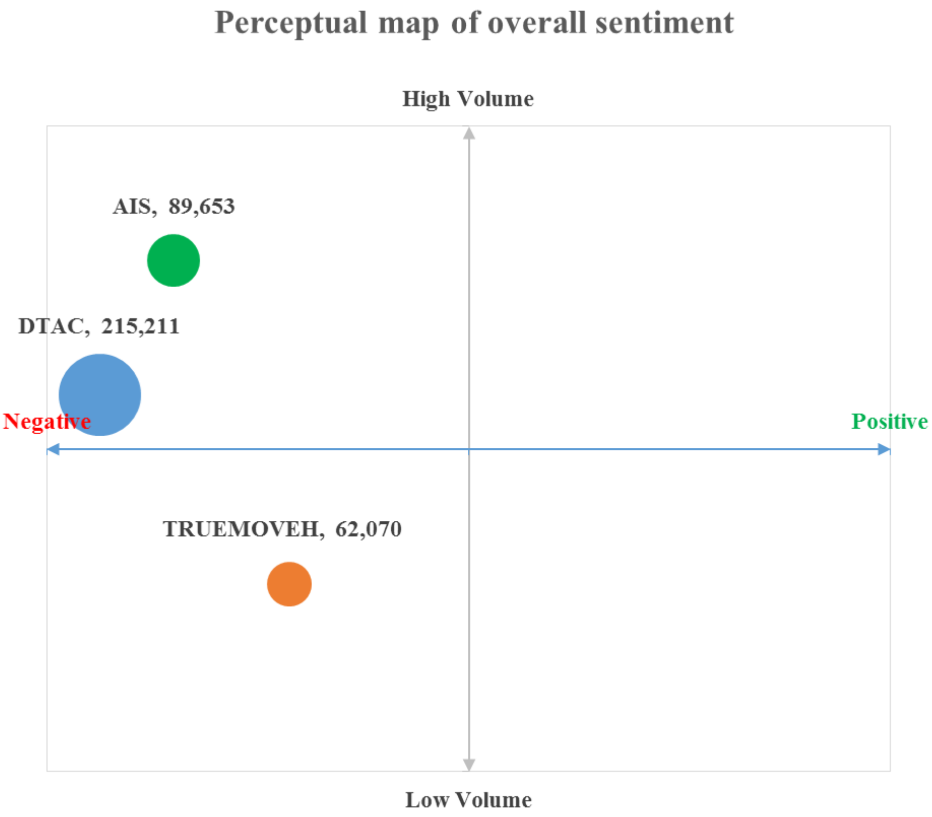


Figure 4.17 Perceptual map of overall sentiment with Tweet volumes

SWOT analysis is key to the CI tool. As mentioned in the Chapter 2 literature review, Strengths (S) are the characteristics of a company that give it an advantage over others and Weaknesses (W) are the characteristics of a company that place the business or project at a disadvantage relative to others. Hence, a radar chart was used to identify the strengths and weaknesses of the three companies by comparing the sentiment of five keyword mentions, as shown in Figure 4.18, which clearly shows the comparison for each category among the three companies.

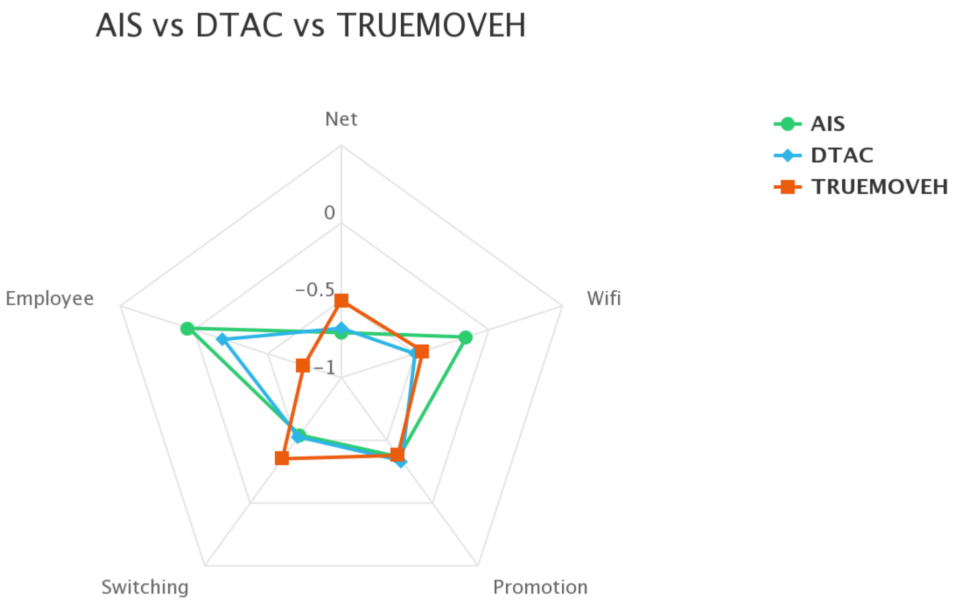
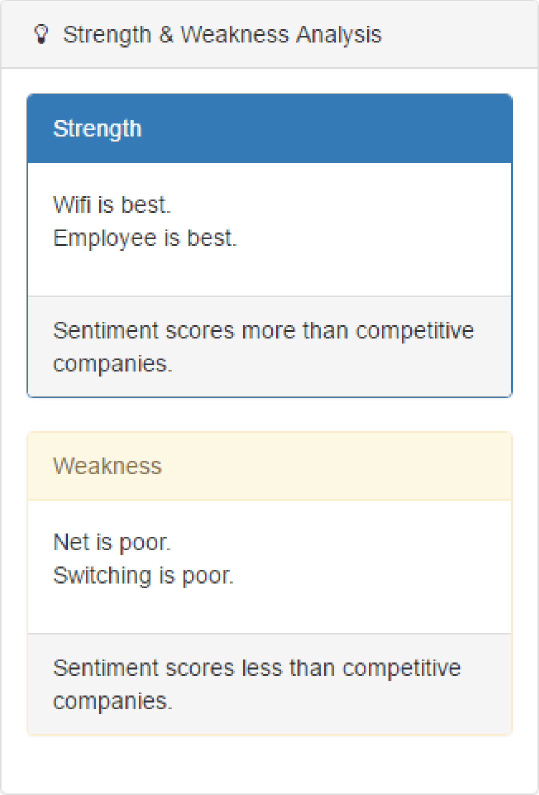


Figure 4.18 Comparison of the overall sentiment for the three companies

For AIS (Figure 4.19), strengths were “employee” and “wifi” and weaknesses were “net” and “switching”; for DTAC (Figure 4.20), a strength was “promotion” and a weakness was “wifi”; and for TRUEMOVEH (Figure 4.21), strengths were “net” and “switching” and weaknesses were “promotion” and “employee”. This illustrates the strengths and weaknesses from customer feedback in social media (Twitter) over the period of six months from October 1, 2014 to March 31, 2015. However, companies could adjust the data collection duration to a shorter time period to monitor the sentiment of customers more closely and realize the strengths and weaknesses of companies in the mindset of its consumers. Furthermore, the CI from social media could become input for strategic or tactical planning for marketing campaigns aimed at improving the weak points of companies.



Strength & Weakness Analysis

Strength

Wifi is best.
Employee is best.

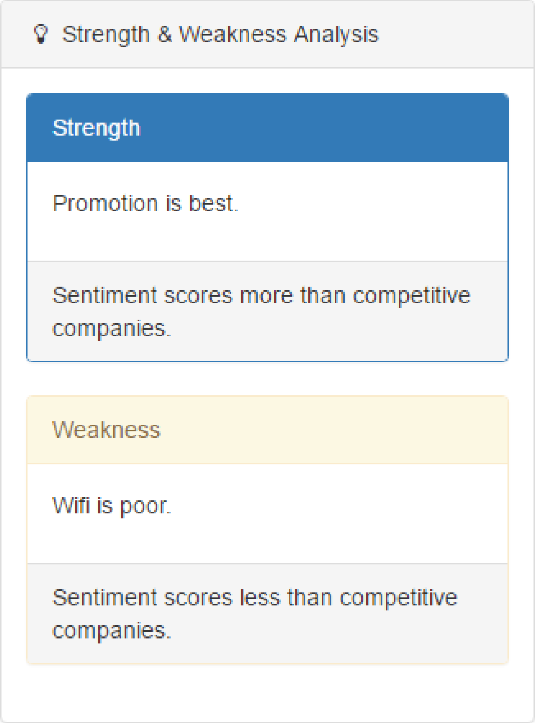
Sentiment scores more than competitive companies.

Weakness

Net is poor.
Switching is poor.

Sentiment scores less than competitive companies.

Figure 4.19 Strength and weakness analysis of AIS



Strength & Weakness Analysis

Strength

Promotion is best.

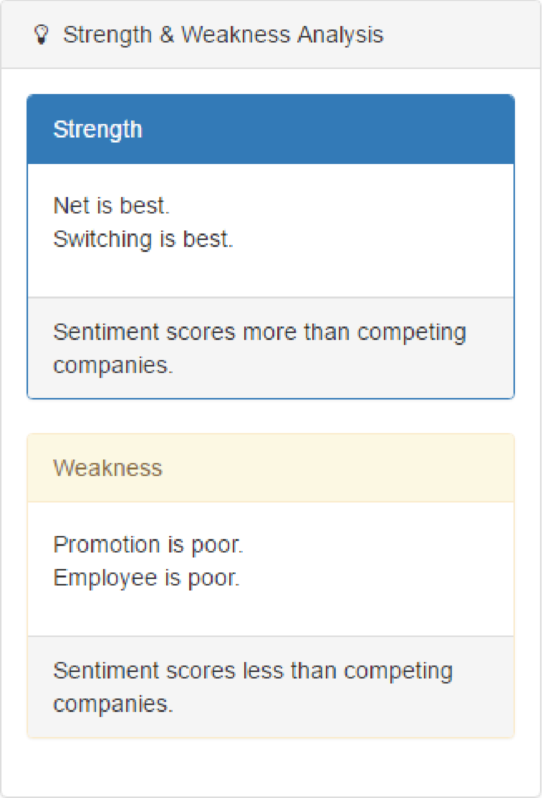
Sentiment scores more than competitive companies.

Weakness

Wifi is poor.

Sentiment scores less than competitive companies.

Figure 4.20 Strength and weakness analysis of DTAC



Strength & Weakness Analysis

Strength

Net is best.
Switching is best.

Sentiment scores more than competing companies.

Weakness

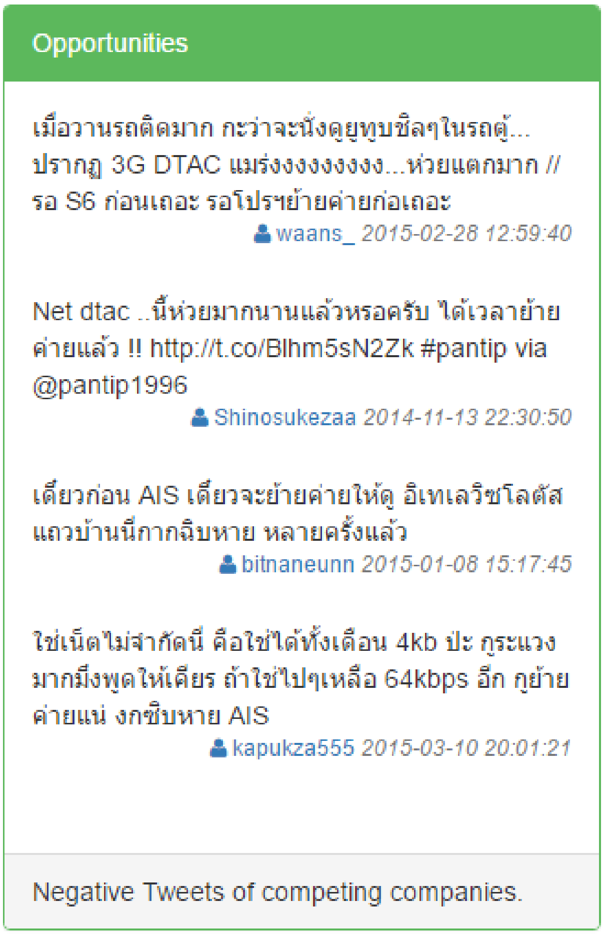
Promotion is poor.
Employee is poor.

Sentiment scores less than competing companies.

Figure 4.21 Strength and weakness analysis of TRUEMOVEH

In addition to strengths and weaknesses, opportunity and threat are also important in SWOT analysis. Opportunities (O) are elements in the environment that a business or project could exploit to its advantage and Threats (T) are elements in the environment that could cause trouble for a business or project. For telecommunications service providers, customer switching is the most pressing issue. This industry in Thailand has become more competitive since 22 September, 2014 when NBTC announced the right of consumers to switch mobile operator without having to change their mobile numbers. So, opportunities for telecommunications service providers are clearly identified when focus on “switching (ย้ายค่าย)”.

Opportunity events aimed at enticing customers away from competing companies by offering those better choices and value appear to have caused negative sentiment Tweets among the competing companies, particularly when mentioning the keyword “switching (ย้ายค่าย)”. These Tweets contained negative sentiment toward the companies, complaints, and statements that the senders wanted to switch to another mobile network provider. Figure 4.22 shows examples from the opportunities screen for TRUEMOVEH.



Opportunities

เมื่อวานรถติดมาก กะว่าจะนั่งดูยูทูบชิลๆในรถตู้... ปรากฏ 3G DTAC แมร่งงงงงงงงงง...ห่วยแตกมาก // รอ S6 ก่อนเถอะ รอโปรฯย้ายค่ายก่อเถอะ

waans_ *2015-02-28 12:59:40*

Net dtac ..นี้ห่วยมากนานแล้วหรอครับ ได้เวลาย้ายค่ายแล้ว !! http://t.co/Blhm5sN2Zk #pantip via @pantip1996

Shinosukezaa *2014-11-13 22:30:50*

เดี๋ยวก่อน AIS เดี๋ยวจะย้ายค่ายให้ดู อิเทเลวิซโลตัส แถวบ้านนี่กากฉิบหาย หลายครั้งแล้ว

bitnaneunn *2015-01-08 15:17:45*

ใช่เน็ตไม่จำกัดนี่ คือใช่ได้ทั้งเดือน 4kb ป่ะ กูระแวงมากมึงพูดให้เคียร ถ้าใช่ไปๆเหลือ 64kbps อีก กูย้ายค่ายแน่ งกซิบหาย AIS

kapukza555 *2015-03-10 20:01:21*

Negative Tweets of competing companies.

Figure 4.22 Example of opportunity events for TRUEMOVEH

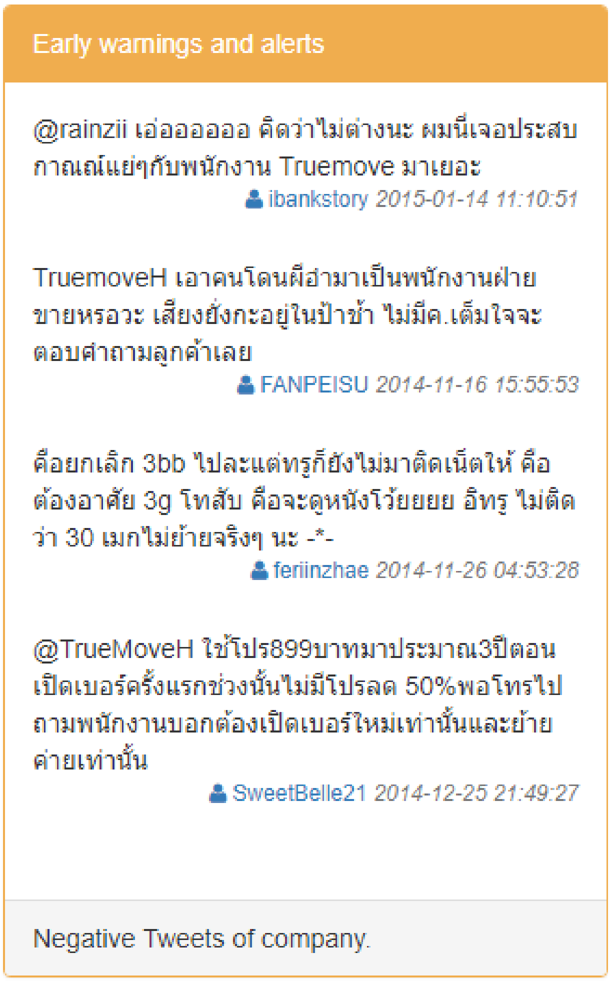
A new enticement Tweet was created with the promotion message related to switching mobile network provider and then using the screen_name field to send this Tweet to competing customers. An example is shown below:

Customer’s Tweet: พรุ่งนี้จะไปย้ายค่ายนะคะ ไม่ต้องมาเทคคงเทคแคร์อะไรกันอีกแล้ว เสื่อมศรัธากับดีแทคมากๆค่ะ @dtac (Tomorrow, I will go to switch mobile network. Don’t take care me again, I lose a lot of confidence with DTAC. @dtac) From screen_name: narinMBA

Enticement Tweet: @narinMBA พิเศษสำหรับคุณ! โปรโมชั่นย้ายค่ายมา TRUEMOVEH ลดราคาทุก Package 50% ดูรายละเอียดเพิ่มเติม http://moveto.truemoveh.com (@narinMBA Special for you! Promotion move to TRUEMOVEH with special discount 50%. More: http://moveto.truemoveh.com)

This is an example of using CI in this study. However, these opportunities require quick action, and so companies should prepare information to create the opportunity to entice customers away from competing companies.

Threats can be analyzed from Tweets posted by competing companies about new services, products, and campaigns, and then by monitoring the feedback of their customers. However, the focus of this research was not on posted Tweets from companies and so this phenomenon was not investigated. However, it was found that negative sentiment Tweets on companies is an early warning sign of vital importance to them. Negative sentiment Tweets on a company are the sign of dissatisfaction in its services, products, or brand, as shown in Figure 4.23.



Early warnings and alerts

@rainzii เอ่ออออออ คิดว่าไม่ต่างนะ ผมนี่เจอประสบกาณณ์แย่ๆกับพนักงาน Truemove มาเยอะ

ibankstory 2015-01-14 11:10:51

TruemoveH เอาคนโดนผีอำมาเป็นพนักงานฝ่ายขายหรอวะ เสียงยังกะอยู่ในป่าช้า ไม่มีค.เต็มใจจะตอบคำถามลูกค้าเลย

FANPEISU 2014-11-16 15:55:53

คือยกเลิก 3bb ไปละแต่ทรูก็ยังไม่มาติดเน็ตให้ คือ ต้องอาศัย 3g โทสับ คือจะดูหนังโว้ยยยย อิทรู ไม่ติดว่า 30 เมกไม่ย้ายจริงๆ นะ -*-

feriinzhae 2014-11-26 04:53:28

@TrueMoveH ใช้โปร899บาทมาประมาณ3ปีตอนเปิดเบอร์ครั้งแรกช่วงนั้นไม่มีโปรลด 50%พอโทรไปถามพนักงานบอกต้องเปิดเบอร์ใหม่เท่านั้นและย้ายค่ายเท่านั้น

SweetBelle21 2014-12-25 21:49:27

Negative Tweets of company.

Figure 4.23 Example of early warnings and alerts for TRUEMOVEH

Companies should pay more attention by monitoring them and quickly responding to any problems to boost customer satisfaction. In particular to this study, negative Tweets that expressed the desire to switch company, an example of which is given below:

Customer’s Tweet: TruemoveH เอาคนโดนผีอำมาเป็นพนักงานฝ่ายขายหรอวะ เสียงยั่งกะอยู่ในป้าช้า ไม่มีค.เต็มใจจะตอบคำถามลูกค้าเลย (TruemoveH. Why do you select employees like zombie? Voice are too bad. They don’t have service mind. They don’t want to answer customers.) From user FANPEISU

FANPEISU expressed negative sentiment on his/her terrible experience in a Tweet, so TRUEMOVEH should have quickly responded and find out the real problem.

In addition to quickly solving an issue, the mention issues in negative sentiment Tweets should be put into the process of complaint management for monitoring and analyzing in depth to create strategic plans to reduce these problems and avoid them happening again.

4.2 Summary

In this chapter, an innovative case study approach for CI in social media was suggested and a process using ASTS (a mixed method of supervised learning techniques and lexicon-based techniques on various companies in any industry) proposed. The results revealed the value of analyzing social media contents by conducting sentiment analysis and benchmarking on the individual top five keyword mention level. The dataset contained Tweets concerning three competing largest mobile operators in Thailand: AIS, DTAC, and TRUEMOVEH. There were 72,661 Tweets in total from a period of six months (October 1, 2014 to March 31, 2015). Tweets not containing the name of one of the three companies were removed, leaving a total of 46,909 Tweets.

The results show that about 47% of the total Tweets (excluded reTweets) were irrelevant and were eliminated by the filtering opinion sub-process. It was therefore necessary to run this sub-process before analyzing customer opinion. Thus, the dataset Tweets were reduced from 46,909 to 24,757, consisting of 11,198 for AIS, 8,435 for

DTAC, and 5,124 for TRUEMOVEH. The sentiment scores for the three companies were negative; the sentiment score for TRUEMOVEH (-0.27) was better than AIS (-0.38) and DTAC (-0.45), but at the keyword mentions level for “employee (พนักงาน)”, the overall sentiment is extremely bad (-0.75).

For CI, benchmarking against competitors comprises essential information. A perceptual map is one of element in CI, as shown in Figure 4.17, which was used to represent the position of customer sentiment for the three companies for the entire period of data collection by measuring Tweet volumes, Tweet sentiment, and Twitter followers. In addition, strength and weakness analysis of the companies was derived using radar charts for the benchmarking of sentiment scores on the top five keyword mentions (net, wifi, promotion, switching, and employee) of the three companies, as shown in Figure 4.18.

Furthermore, examples of using CI in terms of monitoring, opportunity events, and early warning alerts were presented. Opportunity events can help a company to entice customers away from competing companies, a pertinent example of which is shown in Figure 4.22 from the negative sentiment Tweets of the competing companies in this study. A company can create new Tweets that offer a better package than their competitors and send them to the competing companies’ customers. Early warning alerts (Figure 4.23) can be ascertained from negative sentiment Tweets of a company. They are essential for enabling a company to quickly respond to an issue before the negative sentiment concerning it becomes too widespread. In addition, the mention issues in negative sentiment Tweets could be incorporated into the process of complaint management for in depth monitoring and analysis to help create strategic plans to avoid recurrence of this problem.

The another challenges for marketing departments are how to integrate physical channels (offline) and digital channels (online) in a unified customer experience, the Omni-channel. Marketers may send web survey link to ask customers about their experience and level of satisfaction with their products or services, and request more information such as mobile phone numbers, to enable integration of offline and online data. As a result, it will help a business to understand about their customers and deliver a great customer experience to them.

The findings from this study's analysis prove that extracting CI from the analysis of social media content can help businesses to monitor their customers' opinions and compare them with those on their competitors. If a company builds a system to monitor social media for marketing sensing and CI, it can investigate consumers' feedback toward its products and services in real time and can also attempt to entice customers away from competing companies. In addition, the company could combine the analysis CI with internal information and business intelligence (such as sales performance, and campaign and customer behavior tracking) which will give the company an advantage over its competitors.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Summary and Conclusions

The aim of this study was to acquire CI from social media. The main goal was to investigate the possible uses of Twitter information for businesses in Thailand to take advantage of and to solve any associated limitations caused by the semantics of the Thai language. Hence, social media content, specifically Tweets were utilized to generate CI. A new method for Twitter sentiment analysis called ASTS was proposed by using both supervised learning and lexicon-based techniques. Experiments were conducted using Tweets on three mobile network operator companies: AIS, DTAC, and TRUEMOVEH obtained using the Twitter search API focused on Tweets in Thai. A total of 72,661 were collected over a period of six months (from October 1, 2014 to March 31, 2015).

ASTS consists of three modules: (1) data collection, (2) data pre-processing, and (3) classification and evaluation. A collection program was developed to search for keywords in the Twitter feed using the Twitter Search API and setting the language parameter “lang=th” and excluding reTweets. After this, the Tweets were stored in a Tweets databank that acted as input for the data pre-processing module. The process for the data pre-processing module was divided into three steps: (1) Text extraction from the Tweets, (2) Text pre-processing, and (3) Thai word segmentation. For the classification and evaluation module, the main intention was to identify opinion polarity. The classification process was divided into two sub-modules: opinion filtering using supervised learning techniques and opinion polarity identification using lexicon-based techniques.

The results of testing the proposed ASTS method showed significant improvement on the basic concept of using the TOM framework. An average accuracy of 84.80% was achieved, which showed a great improvement (18.67%) over the TOM framework (66.13%). In particular, this clearly shows that opinion filtering helped to analyze Tweets more accurately.

An innovative case study approach for CI in social media aptly demonstrated the use of ASTS. Out of a total of 20,269 Tweets, 9,631 mentioned AIS (47.52%), 7,099 mentioned DTAC (35.02%), and 3,539 mentioned TRUEMOVEH (17.46%). The sentiment scores from the analysis results of using ASTS showed the overall customer sentiment for the companies. The sentiment score for TRUEMOVEH (-0.27) was slightly better than AIS (-0.38) and DTAC (-0.45). Benchmarking against competitors is essential information for CI, and perceptual maps are one of elements in CI. This perceptual map represents the positions of customers' sentiment for the three companies during the entire six-month period of data collection as Tweet volume, Tweet sentiment, and Twitter followers. In addition, strength and weakness analysis on the companies was derived using radar charts of the benchmarking of sentiment scores on the top five keyword mentions (net, wifi, promotion, switching, and employee), as shown in Figure 4.18. In addition, co-occurrence of keyword mentions was used to better understand what customers were discussing.

Furthermore, examples of using CI in terms of monitoring, opportunity events, and early warning alerts were presented. Opportunity events can be advantageous in response to negative sentiment Tweets on competing companies and can help a company to entice customers away from competing companies. Early warning alerts are based on negative sentiment Tweets on a company from which it should quickly identify customer dissatisfaction and then correct the associated problem.

Due to the fact that feedback from social media is direct and expresses the real feeling of users, it is therefore a crucial source of data for VOC. In order to improve customer services and satisfaction, companies should quickly respond to customers to obtain more detail and solve the problem. In addition, the information from social media can give customers knowledge, especially on experience of using products and services. Companies can use users' screen_name to contact and engage with its customers. Marketing departments can send web survey link to ask customers about their experience and level of satisfaction with their products or services, and request more information such as mobile phone numbers, to enable integration of offline and online data, which constitutes the challenge for this era. For integration of physical channels (offline) and digital channels (online) in a unified customer experience, the Omni-channel business model has been developed. That will help a business to understand about their customers and deliver a great customer experience to them. As a result, companies should monitor sentiment using CI over several periods of time such as day, week, and month for different purposes, and subsequently feed this information into a complaint management and decision-making process to solve weakness or plan for new services and products.

In conclusion, the results of this study show the usefulness of the proposed method for theoretical reference and as a practical guide. The findings from this analysis prove that CI extracted from social media content can help businesses to understand their customers' opinions and compare them with those of their competitors. As a result, this research illustrates that CI from analyzing social media content has great potential to produce useful information, actionable knowledge, and critical insights for companies to enhance competitiveness and solve business problems. If companies built systems to monitor social media for VOC sensing and CI, they would be able to investigate consumers' feedback toward their products and services in real time and could entice customers away from their competitors. In addition, companies could combine CI analysis with internal information and business intelligence (such as sales performance, campaign tracking, and customer behavior) which will give them a powerful advantage over their competitors. Moreover, the filtering techniques in this study can be made use of other applications; Tweets that are relevant to a particular company could be useful for various applications such as brand monitoring, campaign monitoring, competitive analysis, and customer engagement.

5.2 Future Research

Although good progress has been made on sentiment analysis in this research, there are still a number of limitations which could lead to many possible directions for future work, such as the analysis of comparative sentences which contain more than one company and the case of a long comment from another social media tools. These are the interesting challenges for researchers in this area. Moreover, a lot of challenges still exist in the social media analytics field; for example, there are still issues with sarcasm and irony. Many people also write spam reviews on social media to promote their own products by giving them positive opinions or to discredit their competitors. As with many languages, a word in Thai contains many meanings depending on the context. However, in Thai, there are no spaces between words and spaces indicate the end of a clause or sentence, and so Thai word segmentation is an important step. Moreover, there are many Thai styles of writing such as mixing Thai and transliterated words in particular contexts and the making of new slang words and emoticons, which has become a new trend for Thais using social media. Thus, improving the accuracy of extracting real customer opinion and sentiment from the social media data is of great interest.

Furthermore, applying this proposed method to other business domains would also be challenging, such as airline, hospital and tourism industries. This proposed framework can be implemented to them by adjusting the related words in ASTS. Another point is to investigate whether there is any correlation between customer sentiment from social media and structured data like sales data and stock prices.

BIBLIOGRAPHY

Albarran, A. B. 2013. **The Social Media Industries.** Abingdon-on-Thames, UK: Routledge.

Alisa Kongthon; Choochart Haruechaiyasak; Chatchawal Sangkeettrakarn; Pornpimon Palingoon and Warunya Wunnasri. 2011. Hotel Opinion: An Opinion Mining System on Hotel Reviews in Thailand. In **Proceedings of Technology Management in the Energy Smart World.** Portland, OR. Pp. 1–6.

Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O. and Passonneau, R. 2011. Sentiment Analysis of Twitter data. In **LSM '11 Proceedings of the Workshop on Languages in Social Media.** PA: Association for Computational Linguistics. Pp. 30–38.

Asian WordNet Project. 2007. **Thai WordNet**. Retrieved 9 January 2016 from http://awn.iisilab.org

Asur, S. and Huberman, B. A. 2010. Predicting the Future with Social Media. In **Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.** Vol. 1. Washington, DC: IEEE Computer Society. Pp. 492–499.

Bakshi, R. K.; Kaur, R.; Kaur N. and Kaur G. 2016. Opinion Mining and Sentiment Analysis. In **Proceedings of the Third International Conference on Computing for Sustainable Global Development (INDIACom).** IEEE Conference Publications. Pp. 452-455.

Bifet, A. and Frank, E. 2010. **Sentiment Knowledge Discovery in Twitter Streaming Data.** Retrieved September 27, 2014 from https://doi.org/10.1007/978-3-642-16184-1_1

Choochart Haruechaiyasak; Alisa Kongthon; Pornpimon Palingoon and Chatchawal Sangkeettrakarn. 2010. Constructing Thai Opinion Mining Resource : A Case Study on Hotel Reviews. In **Proceedings of the Eighth Workshop on Asian Language Resources.** Beijing, China. Pp. 64–71.

Choochart Haruechaiyasak; Alisa Kongthon; Pornpimon Palingoon and Kanokorn Trakultaweekoon. 2013. S-Sense : A Sentiment Analysis Framework for Social Media Sensing. In **Workshop on Natural Language Processing for Social Media.** Nagoya, Japan. Pp. 6–13.

Dai, Y. 2013. **Designing Text Mining-Based Competitive Intelligence Systems.** Doctoral dissertation, University of Eastern Finland. Retrieved August 15, 2015 from http://epublications.uef.fi/pub/urn_isbn_978-952-61-1186-5/urn_isbn_978-952-61-1186-5.pdf

Davenport, T. H. and Harris, J. G. 2007. **Competing on Analytics: The New Science of Winning**. Boston, MA: Harvard Business School Press.

Dey, L.; Haque, S. M.; Khurdiya, A. and Shroff, G. 2011. Acquiring competitive intelligence from social media. In **Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data**. Beijing, China.

Dong, H. 2013. **Social Media Data Analytics Applied To Hurricane Sandy**. Master’s thesis, University of Maryland. Retrieved November 1, 2014 from Proquest Dissertations and Theses Fulltext.

Fan, W. and Gordon, M. D. 2014. The Power of Social Media Analytics. **Communications of the ACM**. 57(6), 74–81.

Gaffney, D. 2010. #iranElection : Quantifying Online Activism. In **Proceedings of the WebSci 10: Extending the Frontiers of Society On-Line**. Raleigh, NC.

Glance, N.; Hurst, M.; Kigam, K.; Siegler, M.; Stockton, R. and Tomokiyo, T. 2005. Deriving Marketing Intelligence from Online Discussion. In **Proceedings from the Eleventh AMV SIGKDD International Conference on Knowledge Discovery in Data Mining**. Chicago, IL. Pp. 419–428. Retrieved October 1, 2014 from http://doi.org/10.1145/1081870.1081919

Goff, D. H. 2013. A History of the Social Media Industries. In **The Social Media Industries.** Albarran, A. B., ed. Abingdon-on-Thames, UK: Routledge. Pp. 21-22.

Goncalves, C. 2014. **Twitter-text.** Retrieved January 9, 2015 from https://github.com/Twitter/Twitter-text

He, W.; Zha, S. and Li, L. 2013. Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry. **International Journal of Information Management.** 33, 464–472. Retrieved September 1, 2014 from http://doi.org/10.1016/j.ijinfomgt.2013.01.001

Java, A.; Song, X.; Finin, T. and Tseng, B. 2007. Why We Twitter : Understanding Microblogging. **Proceedings of the Joint 9th WebKDD and 1st SNA-KDD 2007 Workshop**. San Jose, CA. Pp. 56–65. Retrieved September 1, 2014 from http://doi.org/10.1145/1348549.1348556

Kaplan, A. M. and Haenlein, M. 2010. Users of the World, Unite! The Challenges and Opportunities of Social Media. **Business Horizons.** 53(1), 59–68. Retrieved July 28, 2014 from http://doi.org/10.1016/j.bushor.2009.09.003

Khan, F. H.; Bashir, S. and Qamar, U. 2014. TOM: Twitter Opinion Mining Framework using Hybrid Classification Scheme. **Decision Support Systems.** 57(1): 245–257. Retrieved August 9, 2015 from http://doi.org/10.1016/j.dss.2013.09.004

Kim, Y.; Dwivedi, R.; Zhang, J. and Jeong, S. R. 2016. Competitive Intelligence in Social Media Twitter: iPhone 6 vs. Galaxy S5. **Online Information Review.** 40(1): 42–61. Retrieved May 14, 2016 from https://doi.org/10.1108/OIR-03-2015-0068

Kumar, V. and Reinartz, W. 2012. Relationship Marketing and the Concept of Customer Value. In **Customer Relationship Management**. Heidelberg, Germany: Springer. Pp. 21–31. Retrieved November 3, 2014 from http://doi.org/10.1007/978-3-642-20110-3_2

Liu, B. 2012. **Sentiment Analysis and Opinion Mining**. San Rafael, CA: Morgan & Claypool Publishers. Retrieved September 13, 2014 from https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

Merlin, S.; Bond, A. and Foss, B. 2004. **Consumer Insight: How to Use Data and Market Research to Get Closer to Your Customer**. London, UK: Kogan Page Publishers.

Miner, G.; Elder, J.; Fast, A.; Hill, T.; Nisbet, R. and Delen, D. 2012. **Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.** Waltham, MA: Academic Press.

Mogollón, I. E. 2014. **Multilingual Use of Twitter: Language Choice and Language Bridges in a Social Network.** Doctoral dissertation, The University of Maryland. Retrieved July 12, 2014 from Proquest Dissertations and Theses Fulltext.

NECTEC. 1994. **LexTo - Thai Lexeme Tokenizer (in Thai).** Retrieved August 9, 2014 from http://www.sansarn.com/lexto

Orathai Chinakarapong. 2014. Conceptual Metaphor of Thai Curse Words. **Journal of Humanities**. 11(2): 57-76.

Paniagua, J. and Sapena, J. 2014. Business Performance and Social Media: Love or hate? **Business Horizons.** 57(6): 719–728. Retrieved February 14, 2015 from http://doi.org/10.1016/j.bushor.2014.07.005

Pawoot Pongvitayapanu. 2014. **Thailand and Asia Social Media Data 2014**. Retrieved November 13, 2014 from http://www.slideshare.net/pawoot/for-share-thailand-zocial-award-2014-eng-version

Russell, M. A. 2013. **Mining the Social Web Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.** 2nd ed. Sebastopol, CA: O’Reilly Media.

Saiyai Sakawee. 2014. **Thailand Social Media Stats**. Retrieved October 2, 2014 from https://www.techinasia.com/ thailand-social-media-stats-28-million-facebook-45-million-Twitter-17-million-instagram

SCIP. 1986. **Competitive Intelligence – About SCIP.** Retrieved July 19, 2016 from http://www.scip.org

Tarraf, P. 2005. **Competitive Intelligence and Small Companies: A study of Two Industries**. Master’s thesis, Concordia University. Retrieved October 10, 2014 from http://spectrum.library.concordia.ca/8481/1/MR10337.pdf

Tran, H. V. 2012. **Discovering Entities' Behavior through Mining Twitter**. Doctoral dissertation, The University of Iowa. Retrieved October 10, 2014 from http://ir.uiowa.edu/etd/3545

Twitter. 2014. **Twitter Developer Documentation**. Retrieved August 9, 2014 from https://dev.twitter.com/docs

Warunya Wunnasri; Thanaruk Theeramunkong and Choochart Haruechaiyasak. 2013. Solving Unbalanced Data for Thai Sentiment Analysis. In **Proceedings of the 10th International Joint Conference on Computer Science and Software Engineering**. Khonkaen, Thailand. Pp. 200–205.

WEKA. 2014a. **Data Mining Software in Java**. Retrieved January 15, 2015 from http://www.cs.waikato.ac.nz/ml/weka

WEKA. 2014b. **Text categorization with WEKA**. Retrieved January 30, 2015 from https://weka.wikispaces.com/Text+categorization+with+WEKA

Wikipedia. 2014a. **Twitter.** Retrieved April 20, 2014 from https://en.wikipedia.org/wiki/Twitter

Wikipedia. 2014b. **Voice of the Customer.** Retrieved November 9, 2014 from http://en.wikipedia.org/wiki/Voice_of_the_customer

Wikipedia. 2015a. **List of Emoticons.** Retrieved January 9, 2015 from https://en.wikipedia.org/wiki/List_of_emoticons

Wikipedia. 2015b. **SWOT Analysis.** Retrieved March 29, 2015 from https://en.wikipedia.org/wiki/SWOT_analysis

Wikipedia. 2015c. **Thai Alphabet.** Retrieved January 10, 2016 from https://en.wikipedia.org/wiki/Thai_alphabet

Wikipedia. 2016. **Competitive Intelligence.** Retrieved March 19, 2016 from https://en.wikipedia.org/wiki/Competitive_intelligence

Wiktionary. 2015. **The Free Dictionary**. Retrieved August 9, 2015 from https://th.wiktionary.org

Wilas Chamlertwat. 2011. **Innovative Marketing Tool by Applying Opinion Mining on The Micro-Blog.** Doctoral Dissertation, Chulalongkorn University. Retrieved June 1, 2014 from http://library.car.chula.ac.th/search/Y?search=micro-blog

Yozzo. 2015. **Thailand's Telecom Market Q1 2015**. Retrieved November 25, 2015 from http://www.slideshare.net/yozzo1/thailands-telecom-market-information-q1-2015

Yuan, B. 2016. **Sentiment Analysis of Twitter Data.** Master's thesis, Rensselaer Polytechnic Institute. Retrieved September 11, 2016 from Proquest Dissertations and Theses Fulltext.

Zeng, D.; Chen, H.; Lusch R. and Li, S. 2010. Social Media Analytics and Intelligence. **IEEE Intelligent Systems.** 25(6): 13–16.

Zinner, C. and Zhou, C. 2011. Social Media and the Voice of the Customer. In **The Social Media Management Handbook : everything you need to know to get social media working in your business.** Smith, N.; Wollan, R. and Zhou, C., eds. New Jersey: John Wiley & Sons. Pp. 67–70.

APPENDICES

Appendix A

Word Lists

Table A1 List of abbreviations in text pre-processing step

Abbreviations	Replaced Words	No.
ลค., ลค.	ลูกค้า	2
ตจว., ตจว	ต่างจังหวัด	2
พนง., พนง	พนักงาน	2
บช.,	บัญชี	1
มค.	ม.ค.	1
กพ.	ก.พ.	1
มีค.	มี.ค.	1
เมย.	เม.ย.	1
พค.	พ.ค.	1
มิย.	มิ.ย.	1
กค.	ก.ค.	1
สค.	ส.ค.	1
กย.	ก.ย.	1
ตค.	ต.ค.	1
พย.	พ.ย.	1
ธค.	ธ.ค.	1
Total		19

Table A2 List of transliterated words in text pre-processing step

Transliterated Words	Replaced Words	No.
ais3g	aisthreeg	1
ais4g	aisfourg	1
ais5g	aisfiveg	1
dtac3g	dtacthreeg	1
dtac4g	dtacfourg	1
dtac5g	dtacfiveg	1
true3g	truemovehthreeg	1
true4g	truemovehfourg	1
ture5g	truemovehfiveg	1
พรีวิว	preview	1
6 พลัส, 6พลัส, 6plus	sixplus	3
วีดีโอ, วิดีโอ, video	vdo	3
ห้า จี, ห้าจี, 5 จี, 5จี, 5 g, 5g	fiveg	6
สี่ จี, สี่จี, 4 จี, 4จี, 4 g, 4g	fourg	6
สาม จี, สามจี, 3 จี, 3จี, 3 g, 3g	threeg	6
128gb	onetwoeightgb	1
64gb	sixfourgb	1
16gb	tensixgb	1
32gb	threetwogb	1
คอนเซป, คอนเซบ	concept	2
ไลฟ์สไตล์	lifestyle	1
ไอจี, อินตาแกรม, อินสตาแกรม, อินตาแกม, อินสตาแกรม	instagram	5
ไตรเนท, ไตรเน็ท, ไตรเนต, ไตรเน็ต	trinet	4
ทรูมันนี่	truemoney	1
ไม่มีสัญญาณ	nosignal	1
ฟรี	free	1
เติมเงิน	prepaid	1

Table A2 (Continued)

Transliterated Words	Replaced Words	No.
รายเดือน	postpaid	1
ไม่ชอบ	notlike	1
รูปไม่ขึ้น, รูปมันไม่ขึ้น	notloadimage	2
ไม่จำกัด, ไม่อั้น	unlimit	2
เพย์สบาย	paysbuy	1
ดิจิตอล, ดิจิทัล	digital	2
อัพเดท, อับเดท, อัพเดต, อับเดต	update	4
อัพโหลด, อับโหลด, อัพโหลต, อับโหลต	upload	4
พันทิป, พันทิบ	pantip	2
อัพ	up	1
แอดมิน, แอดมิล	admin	2
จ็อบ	job	1
รีเซ็ท, รีเซ็ต	reset	2
แอบเปิล, แอปเปิล	apple	2
แอพพลิเคชั่น, แอบพลิเคชั่น, แอพ, แอฟ, แอป	application	5
แท็บเล็ต, แท็ปเล็ต	tablet	2
ฮอต, ฮอท	hot	2
อินเทอร์เน็ต, อินเตอร์เน็ต	internet	2
กูเกิ้ล, กูลเกิ้ล, อากู๋	google	3
เดอะมอลล์, เดอะมอล	themall	2
ออนไลน์, ออน	online	2
ช้อป	shop	1
บลูทูธ, บลูทูท, บลูทูด, บลูทูต	bluetooth	4
ฮ็อตสป็อต, ฮอตปอด, ฮอตปอต	hotspot	3
ไวฟาย, ไวไฟ, วายฟาย, วายไฟ, วิฟิ, wi-fi	wifi	6

Table A2 (Continued)

Transliterated Words	Replaced Words	No.
เน๊ต, เน็ต, เนต, เนท, เน็ท, เน็น, เน้ต	net	7
โรมมิ่ง	roaming	1
โมเม้นท์, โมเม้น	moment	2
คอมเม้นท์, คอมเม้น, เม้นท์, เม้น	comment	4
แฮงค์, แฮง	hang	2
ฟันด์โฟลว์	fundflow	1
บุฟเฟ่ต์, บุฟเฟ่	buffet	2
เว็ปไซต์, เว็บไซต์, เว็ป, เว็บ, เวป, ไซต์, ไซท์	website	7
ยูทูป, ยูทู้ป	youtube	2
เลิฟ	love	1
ไลค์	like	1
เซเรเนด	serened	1
ดาต้า, เดต้า	data	2
ดาวน์โหลด, ดาวโหลด	download	2
โหลด	load	1
non-stop	nonstop	1
สเตตัส, เตตัส, ตัส	status	3
รีเฟรช	refresh	1
เฟรช	fresh	1
เฟสบุค, เฟสบุ๊ค, เฟสบุ๊ก, เฟสบุก, เฟส	facebook	5
ไอโฟน	iphone	1
ไอแพด	ipad	1
พลัส	plus	1
สัญญาน, สัญญาณ, สันยาน, สัญญาญ, คลื่น	signal	6
อแดพเตอร์	adapter	1

Table A2 (Continued)

Transliterated Words	Replaced Words	No.
เพลย์ลิสต์	playlist	1
เอ็กซโป	expo	1
เซ่เว่น	seven	1
แมสเสจ, เมสเสจ	message	2
มหิดล	mahidol	1
เอเชียทีค	asiatique	1
ฟร้อนท์	front	1
เซนทรัล	central	1
สติ๊กเกอร์, ติ๊กเกอร์	sticker	2
แบนวิช, แบนวิต, แบนวิด, ช่องสัญญาณ, ช่องสัญญาน	bandwidth	5
แคมเปญจน์, แคมเปญ	campaign	2
call center, คอลเซนเตอร์, คอลเซ็นเตอร์, คอยเซ้นเตอร์, คอลเซนเตอร์, ควยเซ้นเตอร์, คอลเซ็นเตอร์, ควยเซ้นเตอร์, คอยเซนเตอร์, cc	callcenter	10
แอรนดรอย, แอนดรอย	android	2
ซัมซุง	samsung	1
เอดจ์, เอดช์, เอดจ์, เอจ, เอด	edge	5
ย้ายค่าย, เปลี่ยนค่าย, ย้ายมา, ย้ายไป, ย้าย	switching	5
โปรโมชั่น,โปร	promotion	2
ซิคเว่เบรกเก้	sigvebrekke	1
ซิคเว่	sigve	1
โจอี้จัมพ์, โจอี้	joeyjump	2
ไลน์	line	1
รีวิว	review	1
แชท	chat	1
ล็อกอิน	login	1

Table A2 (Continued)

Transliterated Words	Replaced Words	No.
ออฟฟิศ, ออฟฟิส	office	2
ทวิตเตอร์	twitter	1
ทวิต, ทวิด, ทวีต, ทวีด, ทวีท	tweet	5
แบงกิ้ง, แบงค์กิ้ง	banking	2
เรียลไทม์, เรียลไทม์	realtime	2
เรียล	real	1
แบตเตอรี่, แบต	battery	2
พลีท, พลีส	please	2
เวิร์ค	work	1
ไชน่าโมบาย	chinamobile	1
ไชน่า	china	1
โมบายโฟน, โทรศัพท์มือถือ, มือถือ	mobilephone	3
โมบาย	mobile	1
พรีเซนเตอร์	presenter	1
เครือข่าย	network	1
ctw, centralworld, เซนทรัลเวิลด์, เซ็ลเวิร์ล, เซนเวิลด์	centralworld	5
โรบินสัน	robinson	1
ฟีลกู๊ดดดด, ฟีลกู๊ดดด, ฟีลกู๊ดด, ฟีลกู๊ด	feelgood	4
เซอร์ไพรส์	surprise	1
โซเชียล	social	1
แฟนเพจ	fanpage	1
ซีรี่ย์, ซีรี่	serie	2
สมาร์ทโฟน, สมาร์ตโฟน	smartphone	2
โฟน	phone	1
TOT	ทีโอที	1

Table A2 (Continued)

Transliterated Words	Replaced Words	No.
ซิม	sim	1
ฟิกซ์บรอดแบนด์	fixbroadband	1
บรอดแบนด์	broadband	1
แบ็คโบน, แบคโบน	backbone	2
แพ็กเสริม, แพ็คเกจ, แพคเกจ, แพ็กเกจ, แพกเกจ, แพ็ค, แพ็ก	package	7
ไทม์	time	1
Total		279

Table A3 List of slang words in text pre-processing step

Slang Words	Replaced Words	No.
ฟินนนนนนนน, ฟินนนนนนน, ฟินนนนนน, ฟินนน, ฟินน	ฟิน	5
ห่วยแตกกกกกก, ห่วยแตกกกกก, ห่วยแตกกกก, ห่วยแตกกก, ห่วยแตกก	ห่วยแตก	5
กากกกกกกกกกกก, กากกกกกกกกกก, กากกกกกกกกก, กากกกกกกกก, กากกกกกกก, กากกกกกก, กากกกกก, กากกกกก, กากกกก, กากกก, กากก	กาก	11
55555+, 5555+, 555+, 55+, 5555555555, 555555555, 55555555, 5555555, 555555, 55555, 5555, 555	hahaha	12
เซ็งงงงงงงงงง, เซ็งงงงงงงงง, เซ็งงงงงงงง, เซ็งงงงงงง, เซ็งงงงงง, เซ็งงงงง, เซ็งงงง, เซ็งงง, เซ็งง, เซงงงงงงงงงง, เซงงงงงงงงง, เซงงงงงงงง, เซงงงงงงง, เซงงงงงง, เซงงงงง, เซงงงง, เซงงง, เซงง, เซง	เซ็ง	19
ไหง, ทำมัย, ทามมาย, ทำมาย, ทามมัย	ทำไม	5
คับ, ฮัฟ, ก้าบ, ก๊าบ, ครัช, คร้าบบ, คร้าบ, ฮ๊าาฟฟ, ฮ๊าฟ	ครับ	9

Table A3 (Continued)

Slang Words	Replaced Words	No.
ปรู๊ด, ปี้ดๆ, ปี้ด, ฟุด, ฟูด, ฝุด, มั้กมากกกก, มั้กมากกก, มั้กมากก, มั้กมาก, มากกกกกก, มากกกกก, มากกกก, มากกก, มากก, ปรู๊ดปร๊าด, สุดสุด, สุดๆ	มาก	18
แสดด, ช่างแม่ม, อีเหี้ยยยยยย, อีเหี้ย, อิเหี้ย, อีควาย, ชิบหาย, ควาย, เหี้ยยยยยยย, เหี้ยยยยยย, เหี้ยยยยย, เหี้ยยยย, เหี้ยยย, เหี้ยย, เหี้ย, ฟัค, ส๊าด, ควย, หี, เชี่ย, แตด, เชี้ย, แม่งงงง, แม่งงง, แม่งง, แม่ง, จังไร, จันไร, ถอก, สาดด, สาสส, สัส, สัด, เห้อ, ห่า , ชิบ, แมร่ง, แม่ม, แม้ง, ส้นตีน, สนตีน, ตีน, อีดอก, อิดอก, ดอกทอง, เห้, ค้วย, ฟวย	คำด่า	48
คริคริ, ขริขริ, มุ้งมิ้ง, ฟรุ้งฟริ้ง, ฟิน, ชิวๆ, มุ๊งมิ๊ง, chill	Echappy	8
เลอค่า, เริ่ด, เลิศ	เยี่ยม	3
แป๊ป, แป๊ปนุง, แปป	สักครู่	3
กระดึ๊บ, อืด	ช้า	2
อัลไล, อะรัย	อะไร	2
จำเริญๆ, จำเริญ	เจริญ	2
หงายเงิบ, เงิบ	อึ้ง	2
ปังมาก	ดีมาก	1
เร้ว	เร็ว	1
เทอ	เธอ	1
หนาย	ไหน	1
หนม	ขนม	1
Total		217

Table A4 List of misspelling words in text pre-processing step

Misspelling Words	Replaced Words	No.
โทสับ,โทรศัพย์,โทรสับ,โทรสัพ,โทสัพ,โทรศัพ	โทรศัพท์	6
ใบเส็ด, ใบเสด, ใบเส็จ	ใบเสร็จ	3
สเถียน, สเถียร	เสถียร	2
ลาก่อย, ลาก่อนน	ลาก่อน	2
วิดวะ	วิศวะ	1
สงสาน	สงสาร	1
ก้อ	ก็	1
เห้ย	เฮ้ย	1
Total		17

Appendix B

Data Tables

Table B1 The number of AIS opinion Tweets for a week

Week	Positive	Negative	Neutral	Total
WK1/10	59	194	130	383
WK2/10	68	170	164	402
WK3/10	58	104	94	256
WK4/10	27	72	79	178
WK1/11	64	115	110	289
WK2/11	82	171	173	426
WK3/11	83	148	158	389
WK4/11	97	146	182	425
WK1/12	80	136	143	359
WK2/12	81	138	138	357
WK3/12	77	161	112	350
WK4/12	47	97	81	225
WK5/12	69	152	126	347
WK1/01	79	167	148	394
WK2/01	62	139	109	310
WK3/01	51	104	83	238
WK4/01	49	93	106	248
WK1/02	55	130	111	296
WK2/02	83	135	134	352
WK3/02	88	163	96	347
WK4/02	97	251	141	489
WK1/03	83	216	137	436
WK2/03	80	256	161	497
WK3/03	82	288	151	521
WK4/03	103	312	198	613
Total	1,804	4,058	3,265	9,127

Table B2 The number of DTAC opinion Tweets for a week

Week	Positive	Negative	Neutral	Total
WK1/10	43	124	77	244
WK2/10	81	163	122	366
WK3/10	32	70	67	169
WK4/10	4	6	7	17
WK1/11	27	53	55	135
WK2/11	43	102	97	242
WK3/11	45	106	97	248
WK4/11	41	116	86	243
WK1/12	46	120	96	262
WK2/12	64	116	135	315
WK3/12	60	122	102	284
WK4/12	41	94	101	236
WK5/12	42	238	116	396
WK1/01	57	135	100	292
WK2/01	31	138	87	256
WK3/01	26	75	61	162
WK4/01	37	144	95	276
WK1/02	39	93	100	232
WK2/02	55	128	126	309
WK3/02	53	152	120	325
WK4/02	70	145	138	353
WK1/03	58	138	129	325
WK2/03	58	169	133	360
WK3/03	63	172	154	389
WK4/03	64	151	126	341
Total	1,180	3,070	2,527	6,777

Table B3 The number of TRUEMOVEH opinion Tweets for a week

Week	Positive	Negative	Neutral	Total
WK1/10	26	44	39	109
WK2/10	92	95	91	278
WK3/10	10	34	27	71
WK4/10	7	14	11	32
WK1/11	7	23	22	52
WK2/11	29	69	52	150
WK3/11	28	54	38	120
WK4/11	34	47	40	121
WK1/12	35	52	38	125
WK2/12	27	51	58	136
WK3/12	53	67	57	177
WK4/12	33	60	57	150
WK5/12	37	41	20	98
WK1/01	42	56	59	157
WK2/01	48	45	61	154
WK3/01	24	40	23	87
WK4/01	16	70	51	137
WK1/02	16	63	29	108
WK2/02	38	57	56	151
WK3/02	32	46	54	132
WK4/02	43	64	94	201
WK1/03	30	111	67	208
WK2/03	39	61	50	150
WK3/03	25	74	47	146
WK4/03	28	60	36	124
Total	799	1,398	1,177	3,374

Table B4 The number of AIS opinion Tweets for a month

Month	Positive		Negative		Neutral		Total	Overall Sentiment
2014-10	252	(17.62%)	648	(45.31%)	530	(37.06%)	1,430	-0.44
2014-11	338	(21.30%)	604	(38.06%)	645	(40.64%)	1,587	-0.28
2014-12	315	(21.89%)	588	(40.86%)	536	(37.25%)	1,439	-0.30
2015-01	268	(20.06%)	577	(43.19%)	491	(36.75%)	1,336	-0.37
2015-02	323	(21.77%)	679	(45.75%)	482	(32.48%)	1,484	-0.36
2015-03	414	(17.58%)	1,194	(50.70%)	747	(31.72%)	2,355	-0.49
Total	1,910	(19.83%)	4,290	(44.54%)	3,431	(35.62%)	9,631	-0.38

Table B5 The number of DTAC opinion Tweets for a month

Month	Positive		Negative		Neutral		Total	Overall Sentiment
2014-10	193	(20.27%)	447	(46.95%)	312	(32.77%)	952	-0.40
2014-11	158	(17.67%)	389	(43.51%)	347	(38.81%)	894	-0.42
2014-12	236	(19.02%)	516	(41.58%)	489	(39.40%)	1,241	-0.37
2015-01	166	(13.70%)	654	(53.96%)	392	(32.34%)	1,212	-0.60
2015-02	217	(17.80%)	518	(42.49%)	484	(39.70%)	1,219	-0.41
2015-03	264	(16.70%)	701	(44.34%)	616	(38.96%)	1,581	-0.45
Total	1,234	(17.38%)	3,225	(45.43%)	2,640	(37.19%)	7,099	-0.45

Table B6 The number of TRUEMOVEH opinion Tweets for a month

Month	Positive		Negative		Neutral		Total	Overall Sentiment
2014-10	155	(27.19%)	213	(37.37%)	202	(35.44%)	570	-0.16
2014-11	107	(22.96%)	200	(42.92%)	159	(34.12%)	466	-0.30
2014-12	158	(25.73%)	243	(39.58%)	213	(34.69%)	614	-0.21
2015-01	148	(25.34%)	232	(39.73%)	204	(34.93%)	584	-0.22
2015-02	129	(21.79%)	230	(38.85%)	233	(39.36%)	592	-0.28
2015-03	140	(19.64%)	349	(48.95%)	224	(31.42%)	713	-0.43
Total	837	(23.65%)	1,467	(41.45%)	1,235	(34.90%)	3,539	-0.27

BIOGRAPHY

NAME Jitrlada Rojratanavijit

ACADEMIC BACKGROUND Bachelor Degree in Computer Engineering from Kasetsart University, Bangkok, Thailand in 1994.
Master Degree in Computer Science from Mahidol University,
Bangkok, Thailand in 2002.

PRESENT POSITION Data Processing Officer 10
Information Technology
Development Department
Metropolitan Electricity Authority,
Bangkok, Thailand.