S-Sense: A Sentiment Analysis Framework for Social Media Monitoring Applications

Choochart Haruechaiyasak*, Alisa Kongthon*,
Pornpimon Palingoon*, and Kanokorn Trakultaweekoon*

Abstract

Today the amount of social media usage has increased exponentially. Many businesses and organizations including market research agencies are seeking for tools which could perform real-time sentiment analysis on this explosive "big data" contents. In this paper, we propose S-Sense, a framework for analyzing sentiment on Thai social media. The proposed framework consists of analysis modules and language resources. Two main analysis modules, intention and sentiment, are based on classification algorithm to automatically assign appropriate intention and sentiment class labels for a given text. To train classification models, language resources, i.e., corpus and lexicon, are needed. Corpus consists of a collection of texts manually labeled with appropriate intention and sentiment classes. Lexicon consists of both general terms from dictionary and clue terms which help identifying the intention and sentiment. To evaluate performance and robustness of the analysis modules, we prepare a data set from Twitter posts and Pantip web board in mobile service domain. The experiments are set up to compare the performance between two different lexicon sets, i.e., general and clue terms. The results show that incorporating clue terms into feature vectors for constructing the classification models yield significant improvement in terms of accuracy. The proposed S-Sense framework could be potentially applied for many applications including social media monitoring for improving market brand and campaign management.

Keywords: Sentiment Analysis, Intention Analysis, Social media Monitoring.

1. Introduction

Due to the enormous volume, social media has become recognized as a good example of "big data" contents. One of the challenging issues in handling big data is to perform real-time analysis on the contents. Today social media has been widely accepted as an active communication channel between companies and customers. Many companies regularly use social networking websites to promote new products and services, and post announcements to the customers. On the other hand, customers often post their comments to express some sentiments towards products and services. Many customers also post questions and requests to get answers and helps from the customer services. Due to the real-time nature of the social media, monitoring customers' comments has become a critical task in customer relation management (CRM). Sentiment analysis has received much attention among market research community as an effective approach for analyzing social media contents. Some highlighted applications of sentiment analysis include brand monitoring, campaign monitoring and competitive analysis.

Thailand is among the top countries having a large population on social networking websites such as Facebook and Twitter. Many companies in Thailand start to see the importance of using social media analysis to gain some insight on what people think about their brands, products and

* *Speech and Audio Technology Laboratory (SPT) National Electronics and Computer Technology Center (NECTEC) National Science and Technology Development Agency (NSTDA) Thailand Science Park, Klong Luang, Pathumthani , Thailand.*

services. Although many commercial software tools for social media analysis are available, they do not support Thai language. In this paper, we propose S-Sense, a framework for analyzing sentiment on Thai social media contents. To provide a complete solution, our proposed framework consists of many components including tagging tool, language resources, analysis and visualizing modules.

Among all of the components in S-Sense, language resources are considered very essential for providing the infrastructure to train both intention and sentiment analysis models. In our proposed framework, language resources consist of two components, corpus and lexicon. Corpus consists of a collection of texts manually labeled with appropriate intention and sentiment classes. Lexicon consists of two types of terms, general and clue. The general lexicon includes terms found in LEXiTRON which is a well-known Thai-English electronic dictionary. In S-Sense, the general lexicon is modified by including new terms such as slangs, chat language, transliterated words, found in Thai Twitter corpus. The second lexicon consists of clue terms which help identifying the intention and sentiment. Example of clue terms for sentiment analysis are polar terms (such as “stylish”, “beautiful” and “expensive”), which contain either positive or negative sentiment.

For the analysis modules, we apply classification algorithm to automatically assign appropriate intention and sentiment class labels for a given text. The performance of classification models generally depends on the choice of classification algorithms including parameter settings, the size of training corpus and the design of term feature sets. The current version of S-Sense applies the multinomial Naive Bayes algorithm. The reason we used Naive Bayes is its requirement of a small amount of training data to estimate the parameters for learning the models. Also Naive Bayes is a descriptive and probabilistic machine learning, therefore, the results could be easily analyzed and explained. The classification results are returned with a probability value which could be interpreted as the confidence level. In addition to the proposed framework, another contribution of this paper is the comparative study of using different lexicon sets for training the analysis models. We compare the performance of intention and sentiment analysis models by using two different sets of lexicons, general and clue terms. The evaluation corpus consist of Twitter posts and Pantip web board topics in mobile service domain. The experimental results will be presented along with the discussion on the error analysis.

The remainder of this paper is organized as follows. In next section, we review some related works on sentiment analysis and many different approaches for constructing language resources for sentiment analysis. In Section 3, we present the proposed S-Sense framework for Thai intention and sentiment analysis. Details on each components are given with illustration. Section 4 gives details in text tokenization and normalization for Thai written texts. In Section 5, we evaluate the S-Sense framework by using a data set collected from Twitter and Pantip Thai web board. Examples of difficult cases are discussed along with some possible solutions. In Section 6, we gives examples of potential applications for S-Sense including brand monitoring and social media monitoring. Section 7 concludes the paper with the future work.

2. Related work

Due to its potential and useful applications, opinion mining and sentiment analysis has gained a lot of interest in text mining and NLP communities [1], [2], [3]. Much work in this area focused on evaluating reviews as being positive or negative either at the document level [4], [5] or sentence level [6], [7]. For instance, given some reviews of a product, the system classifies them into positive or negative reviews. No specific details or features are identified about what customers like or dislike. To obtain such details, a feature-based opinion mining approach has been proposed [8].

The problem of developing subjectivity lexicons for training and testing sentiment classifiers has recently attracted some attention. Although most of the reference

corpora has been focused on English language, work on other languages is growing as well. Ku and Chen [9] proposed the bag-of-characters approach to determine sentiment words in Chinese. This approach calculates the observation probabilities of characters from a set of seed sentiment words first, then dynamically expands the set and adjusts their probabilities. Later in 2009, Ku et al. [10], extended their bag-of-characters approach by including morphological structures and syntactic structures between sentence segment. Their experiments showed better performance of word polarity detection and opinion sentence extraction. Haruechaiyasak et al. [11], proposed a framework for constructing Thai language resource for feature-based opinion mining. The proposed approach for extracting features and polar words is based on syntactic pattern analysis.

Our main contribution in this paper is the proposed framework for analyzing intention and sentiment from social media texts. We initially performed some evaluation on Thai texts to show the effectiveness of the proposed components and modules. The proposed framework can be easily extended to support other languages, especially for unsegmented languages, by providing the plugged-in resources including lexicon and corpus.

3. Proposed framework

In this paper, we focus on both language resources and the analysis modules as a complete framework for Thai-language intention and sentiment analysis. The proposed framework could easily be extended to support other languages by constructing language-specific resources. Our framework is also designed for easy adaptation to businesses in different domains. Similar to language-specific support, to apply the proposed framework for a specific domain, one can use the provided tagging tool to prepare domain-specific resources, i.e., annotated corpus and lexicon.

3.1 Components and modules

The proposed S-Sense framework (shown in Figure 1) consists of the following components.

Text collecting & processing: This component involves the process of crawling and collecting social media contents from different websites. The process includes basic text processing, i.e., tokenization and normalization. Term normalization is the process of converting a word as appeared in the text into a predefined term and cleaning extra repeated characters which are not part of the term. For example, a word "thnxsss" can be normalized to the term "thank".

UREKA: The main task of UREKA (Utilization on REsource for Knowledge Acquisition) is to extract key feature terms or phrases from a given text. Terms or phrases which are statistically significant in the corpus can be presented as interesting issues to the users. Another task is to filter and classify a given text into a topic. When collecting texts from social networking websites, it is very common to see many collected texts are not relevant to the brands or products being monitored. Therefore, a classification model could be trained to filter out the irrelevant texts from the corpus. After obtaining the relevant texts, another classification model could be trained to classify each text into a predefined set of topics. For example, in mobile service domain, topics could include signal quality, promotion and customer service.

S-Sense: This is the main analysis component under the framework. S-Sense consists of two analysis modules. Intention analysis classifies each text into four classes: announcement, request, question and sentiment. Sentiment analysis further classifies each text based on its sentiment, i.e., positive or negative. Other components of S-Sense include visualizing modules including adaptive emoticon and interactive dashboard. These modules are used for displaying the summarized reports for the analyzed texts.

Tagging tool and language resources: Under the proposed framework, language resources include two components, annotated corpus with domain and language-specific lexicons. To construct language resources, we provide a tagging tool for linguists to work with. The tagging tool is a web-based application which consists of a DBMS and a GUI.

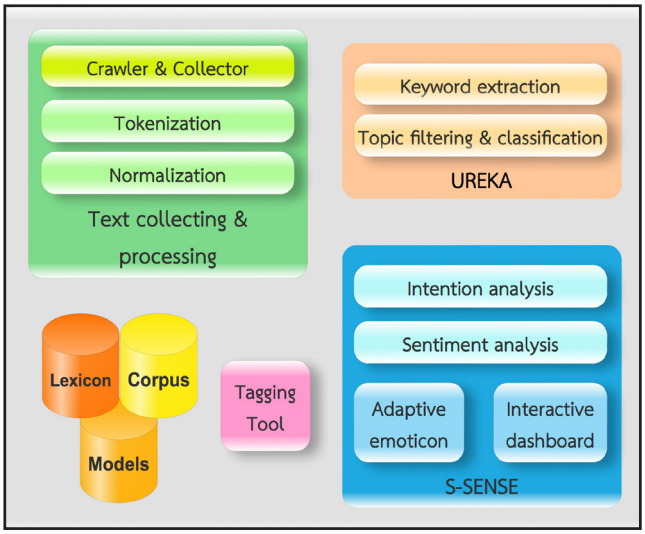

Figure 1. *The proposed S-Sense framework.*

3.2 Analysis tasks

The current version of S-Sense framework focuses on two main analysis modules, intention and sentiment. The intention analysis include the following categories.

Announcement: This type of intention refers to messages or posts in which a company intends to communicate with their customers, e.g., advertisement of new products or event announcement.

Request: This intention is used for customers to ask for help when having trouble or problem with the company's products or services. Customers would expect immediate response from the company to solve the problem.

Question: This intention refers to messages or posts from customers asking for information related to products and services. The question is, for example, a customer's post asking for more details of a new mobile service promotion.

Table 1. *Example of texts categorized by different intentions.*

Intention	Example
Announcement	อัตราค่าบริการ Happy Bonus ปรับปรุงใหม่จ๊ะ เริ่มใช้วันที่ 1 ค่ะ The new service fee for Happy Bonus will start on the 1st of this month.
Request	สมัครใช้บริการ Call Screening เองไม่ได้ CC ช่วยด้วยครับ I can't apply for Call Screening myself. CC (Call Center), please help me.
Question	โปรไหนของ one-2-call ที่รอรับสายได้นานสุดครับ Which promotion package of one-2-call allows the longest call waiting time?
Sentiment	น่ารำคาญมาก DTAC เมื่อไหร่จะปรับปรุงสัญญาณสักที โดยเฉพาะบนBTS Very annoyed. DTAC, when will you improve the signal? Especially on the BTS.

Sentiment: This intention is when customers express their opinions or sentiments towards the company's brand, products and services. Sentiment can be divided into positive, neutral and negative aspects.

Table 1 shows some examples of sentences categorized in each intention. It is important to analyze intention before performing sentiment analysis. Without intention analysis, a sentence containing positive polar words such as an advertisement would be identified as containing the sentiment intention. For example, a sentence "The new high-speed Internet is faster and cheaper. Apply today at the shop near you." is an advertisement, but could be incorrectly identified as having positive sentiment. Therefore, Identifying a sentence as announcement or advertisement would help improve overall performance of sentiment analysis.

4. Text Tokenization and Normalization

Social media texts include chat message, SMS, comments and posts. These texts are usually short and noisy, i.e., contain some ill-formed, out-of-vocabulary, abbreviated, transliterated and homophonic transformed terms. These special characteristics are due to many reasons including inconvenience in typing on virtual keyboards of smartphones and intentional transformation of existing terms to better express the emotion and feeling of the writers. As a result, performing basic text processing tasks such as term tokenization has become much more challenging. Many previous works in text normalization were proposed to handle English texts such as SMS and social media [12 - 17].

Tokenizing Thai written texts is more difficult than other languages in which word boundary markers are placed between words. Thai language is considered as an unsegmented language in which words are written continuously without the use of word delimiters. Word segmentation is considered a basic yet very important NLP task in many unsegmented languages. The main goal of word segmentation task is to assign correct word boundaries on given text strings.

From our preliminary study, the errors from tokenizing Thai texts from social media texts are due to four cases: insertion, transformation, transliteration and onomatopoeia. These error types are considered as intentional errors, which are caused by users intentionally create, alter and transform existing words on different purposes.

Table 2. *Intentional spelling error types and examples.*

Intentional Spelling error	Example
(1) Insertion	มากกกกก (มาก = very) โอ๊ยยยยย (โอ๊ย = ouch!) ว้าววววว (ว้าว = wow) แล้ววววว (แล้ว = already)
(2) Transformation (2.1) Homophonic	มั่ก (มาก = very) เด่ว (เดี๋ยว = just) ร้าก (รัก = love) อะเคร (โอเค = okay) เมิง (มึง = you) กรู (กู = I) e-ngo (อีโง่ = idiot) จุงเบย (จังเลย = really) 555 (ฮ่าๆๆ = ha ha ha)
(2.2) Syllable trimming	มหาลัย (มหาวิทยาลัย = university) โอ (โอเค = okay) มอไซค์ (มอเตอร์ไซค์ = motorcycle) เตี๋ยว (ก๋วยเตี๋ยว = noodle)
(3) Transliteration	นอย (พารานอยด์ = paranoid) ชิว ชิว (chill chill)
(4) Onomatopoeia	แง (baby crying sound) ตุ้ม (explosion sound) จุ๊บ (kissing sound) เอี๊ยด (braking sound)

Table 2 lists all error types with some examples. The original terms or brief descriptions are shown in parentheses with translations. Each error type is fully explained as follows.

Insertion: This type of error is caused by repeated characters at the end of a word. This error type also appears in English, e.g., whatttt, sleepyyy and loveeee.

Transformation: This error type is caused by transformation of existing terms and can be categorized into two following types.

Homophonic: The homophonic terms refer to terms with the same or similar pronunciation. A homophonic term is normally created by replacing an original vowel with a new vowel which has similar sound. Some examples in English are luv (love), kinda (kind of) and gal (girl).

Syllable trimming: The syllable trimming is a transformed term by deleting one or more syllables from an existing term for the purpose of reducing the keystrokes.

Transliteration: Thai Transliterated terms are newly created terms converted from other language scripts. Transliterated terms are commonly found in modern Thai written texts, e.g., chat and social media. Most of the terms are transliterated from English terms including named entities such as company and product names.

Onomatopoeia: Onomatopoeia terms are created by using characters to form new terms to imitate different sounds in nature and environment including humans and animals [18], [19], [20]. Onomatopoeia terms are typically used in chat and social media texts to make the communications between users more vivid. For example, to make the kissing action sound more realistic, the word joob in Thai (or smooch in English) which imitates the kissing sound is normally used.

In our previous work [21], we proposed LexToPlus, an algorithm for tokenizing and normalizing Thai texts. The proposed algorithm performs tokenization with normalization process. Based on the experiment on Twitter corpus, our proposed algorithm achieved the overall accuracy equal to 96.3%. We apply LexToPlus for performing tokenization and normalization processes in the S-Sense framework

5. Experiments and discussion

To evaluate the proposed S-Sense framework, we perform experiments using a corpus in the domain of mobile service. The corpus is obtained between March and June in 2013 from two sources, Twitter and Pantip , one of the top visited web boards in Thailand. The total number of randomly selected texts in the corpus is 2,723. The corpus was annotated in two aspects, intention and sentiment. Table 3 summarizes the number of tagged texts in four different intentions. The majority of intentions is sentiment which accounts for approximately 64% of the corpus. The reason is when using social networking websites or web boards, users often express their opinion and sentiment more than other intentions.

Table 3. *Number of tagged texts in four different intentions.*

Intention	# Texts
Announcement	94
Request	405
Question	456
Sentiment	1,768
Total	**2,723**

Table 4. *Number of tagged texts in positive and negative sentiments.*

Sentiment	# Texts
Positive	156
Negative	1,612
Total	**1,768**

For the sentiment intention, we further annotated each text based on its sentiment, i.e., positive or negative. Table 4 summarizes the number of tagged texts in positive and negative sentiment. It can be observed that negative sentiment accounts for approximately 91%. This is not very surprising since users tend to complain when having problems using the mobile service. Major reported problems in mobile service industry include, for example, weak or unavailable signal, call drop, slow data transfer rate, impolite service and long waiting time for call center.

Table 5. *Example of annotated texts categorized by different intentions and sentiments.*

Intention		Example
Announcement		อัตราค่าบริการ Happy Bonus ปรับปรุงใหม่จ๊ะ เริ่มใช้วันที่ 1 ค่ะ The new service fee for Happy Bonus will start on the 1st of this month.
		โปรใหม่!! ทรูมูฟ... ชิมสุดคุ้ม โปรวินาทีละ 1 ส.ต. ตลอด 24 ชั่วโมง New promotion!! True Move... Best-deal SIM, 1 satang / second all day and night.
Request		สมัครใช้บริการ Call Screening เองไม่ได้ CC ช่วยด้วยครับ I can't apply for Call Screening myself. CC (Call Center), please help me.
		รบกวนCC AISหน่อยค่ะ..เงินในโทรศัพท์หายไปไหนไม่รู้ (- -)?? AIS Call Center, please.. My pre-paid balance has gone missing without a clue ??
Question		โทรศัพท์หาย จะทำซิมใหม่เบอร์เดิมของ ais ต้องใช้เอกสารอะไรบ้างครับ I lost my phone. To get a new SIM card, what documents are required?
		โปรไหนของ one-2-call ที่รอรับสายได้นานสุดครับ Which promotion package of one-2-call allows the longest call waiting time?
Sentiment	Negative	น่ารำคาญมาก DTAC เมื่อไหร่จะปรับปรุงสัญญาณสักที โดยเฉพาะบนBTS Very annoyed. DTAC, when will you improve the signal? Especially on the BTS.
	Positive	ขอบคุณและชื่นชม เจ้าหน้าที่ AIS serenade call center ประทับใจครับ Thank you to the operator at AIS serenade call center. Very impressive.

Table 5 shows some examples of annotated corpus in different intention and sentiment. In addition to annotating each text with an intention label, we collect clue terms which could help identify the intention. For example, from the announcement intention, the terms and phrases "new promotion", "best-deal" and "will start on" are collected into the clue lexicon. From the sentiment intention, we collected the terms "annoyed" and "impressive". Other clue terms are underlined for each example in the table.

Table 6. *Two types of lexicons: general and clue.*

Lexicon		# Terms
General	Lexitron	35,328
	Twitter	1,341
Clue	Announcement	86
	Request	177
	Question	454
	Polar (Negative)	1,675
	Polar (Positive)	1,237

Table 6 shows the statistics of lexicons used in the experiments. There are two types of lexicons: general and clue terms. General lexicon include two sets of terms, LEXiTRON, which are general words from Thai dictionary, and Twitter which contains newly found words from Thai Twitter corpus. Words obtained from Twitter include slangs and transliterated words from other languages. Clue lexicon include terms or phrases which could help identify intention and sentiment. One of the main objectives in the experiments is to observe the effect of incorporating clue lexicon in constructing classification models for intention and sentiment analysis. Therefore, we perform a comparative study on using different sets of lexicons.

To perform experiments, we apply the multinomial Naive Bayes algorithm to learn the classification models [22]. The reason we use Naive Bayes is due to the small number of sample texts in the corpus, especially for the announcement intention. Naive Bayes only requires a small amount of training data to estimate the parameters for learning the models. Also Naive Bayes is a descriptive and probabilistic machine learning, therefore, the results could be easily analyzed and explained. The classification results are returned with a probability value which could be interpreted as the confidence level.

The first experiment is the intention analysis. For each intention, we train a binary classification model with two classes, related and other. If a given text is analyzed as containing a particular intention, it will be assigned with the class label related. We prepare the data set by using the same amount of texts in each class. For example, in announcement intention, we use 94 announcement texts and randomly select

another 94 texts from other intentions. To see the advantage of using clue terms as additional term feature, we compare the results between using only general lexicon and using both general and clue lexicons. The performance metric is accuracy which is defined as the number of correctly classified instances over the total number of test instances.

Table 7. *Experimental results on intention analysis.*

Intention	Term feature	Accuracy (%)
Announcement	General	78.72
	General + Clue	**80.85**
Request	General	63.08
	General + Clue	**69.38**
Question	General	73.13
	General + Clue	**79.82**
Sentiment	General	67.47
	General + Clue	**73.61**

Table 7 shows the experimental results for intention analysis. The results are based on 10-fold cross validation. From the table, it can be observed that adding clue terms into the term feature helps improve the classification accuracy for all intentions. Especially for request, question and sentiment, the improvement is over 6%. For announcement, the improvement is approximately 2%. This is probably due to the difficulty in defining and collecting the clue terms for announcement intention. For example, some of the terms like "new" must be collocated with other term in a phrase, e.g. "new promotion". As the phrase becomes more specific, it will not be found in the test instances. Another observation is the request intention is the most difficult to analyze. This is due to often when users wish to request for something, there is no specific term or clue term in the message. The request intention is implicitly expressed with verbs or polar terms, therefore causing confusion to other intention classes.

The second experiment is the sentiment analysis. We train a binary classification model with two classes, positive and negative. Table 8 shows the experimental results on sentiment analysis. The results are based on 10-fold cross validation.

Table 8. *Experimental results on sentiment analysis.*

Term feature	Accuracy (%)
General	89.55
General + Clue	**91.64**

From the table, we can observe that using clue terms as additional term features helps increase the accuracy by approximately 2%. The small amount in improvement is probably due to terms in general dictionary and from Twitter contain sentiment which already helps identify the polarity of the texts.

To perform error analysis, we look at the test instances which are misclassified, i.e., classifying positive into negative and vice versa. We can summarize two major causes of errors as word sense ambiguity and sarcasm. The first problem occurs when a polar term contains both positive and negative senses depending on the contexts. For example, the word "strong", when appearing with the term "signal" will give positive polarity. However, when it appears with the term "employee", the term has the meaning of "impolite" and a negative polarity should be assigned. However, due to the small corpus size and simple feature vector which treats each term independent, sometimes, the terms cannot be learned properly. To solve this problem, we will explore the idea of incorporating contextual terms with the clue terms in our future work. Each clue term will be associated with some context terms to identify the polarity of the texts.

The second problem is sarcasm which is much more difficult to solve. This problem is still a difficult and challenging task in sentiment analysis of any languages [23], [24], [25]. While there are some research work to identify sarcasm in given texts, the performance is still poor. However, some of the sarcastic texts can still be identified by detecting some common slangs which are usually used in sarcastic texts. In Thai language, if users express a positive sentiment in an exaggerated way or in a contradicting way, then the message is most likely sarcastic. For example, "Today the download speed is faster than the speed of light. Thank you very much!" is considered as sarcastic.

6. Potential Applications

S-Sense can be applied in many different applications. Some of the potential applications are as follows.

Brand monitoring: With the widespread of social media, today customers have more freedom to express their sentiments towards products and services. Analyzing sentiments of the customers could help companies gain some insight on how they feel when using their products and services. More importantly, many companies are highly associated with their brands. Negative sentiments towards the company's brand could have negative impact on the product sales. Therefore, it is very important for companies to monitor or track the mentions and sentiments of the customers on social media.

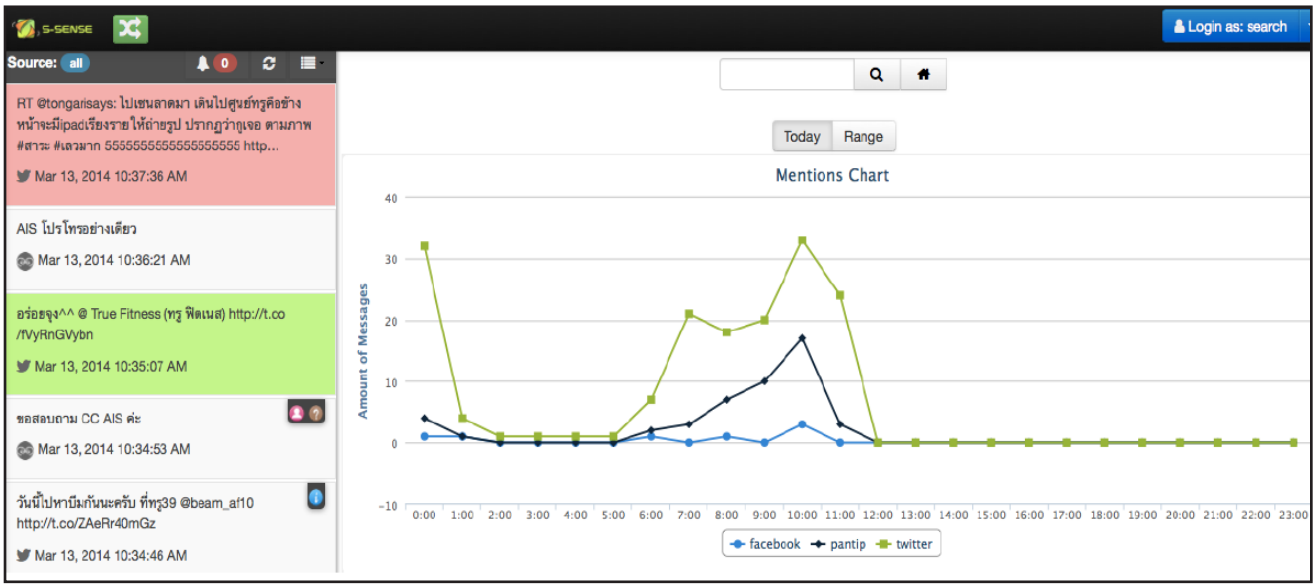

Figure 2. *S-Sense brand monitoring dashboard: mention chart.*

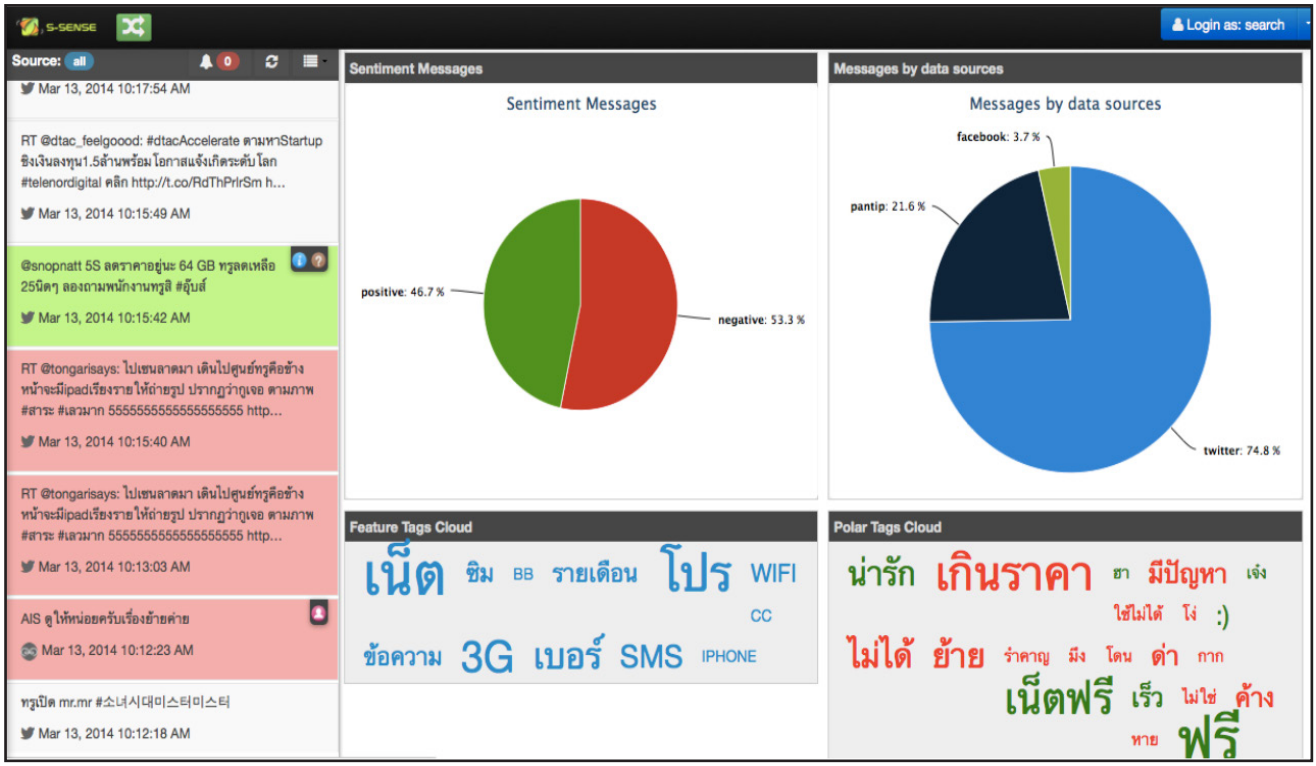

Figure 3. *S-Sense brand monitoring dashboard: sentiment chart and tag clouds.*

Campaign Monitoring: Many times throughout the year, the company would launch different campaigns involving new products and services. The goal of campaign monitoring (i.e, tracking) is to measure the customers' feedback on each campaign. The results could be analyzed in terms of number of mentions, positive and negative sentiments and the key product or service features in which customers feel positive or negative about.

Competitive Analysis: This task is to monitor and analyze the activities including sentiments of customers towards the company's competitors. The analysis results could help gain some insight on strengths and weaknesses of the competitors in the market. For example, if a competitor has many complaints on certain product features, the company could grab the opportunity by advertising its own product features which are better than the competitor's.

Employee Engagement: One of the main problems in many organizations today is the high turnover rate. One of the solutions is to monitor and analyze the employee engagement level. This task is to measure the employees' sentiments towards their jobs, colleagues and organization. The measure could reveal how much employees are willing to learn and perform at work, and to get involved in different activities initiated by the organization.

Figure 2 shows an application of S-Sense for brand monitoring in mobile service business domain. The application is designed for monitoring real-time social media feeds from Facebook, Twitter and Pantip, the most visited webboard in Thailand. On the captured screen, the left panel displays real-time posts updated within 5-10 minutes. Each post is sent to S-Sense engine to analyze for the user's intention and sentiment. For example, a post with message, "Please let me ask CC (call center) from AIS.", is classified as having both request and question intentions and is denoted with a request and question icons on the top right corner of the feed. A post with either positive or negative sentiment is highlighted with green or red color, accordingly. The right panel in Figure 3 is a mention chart which displays the number of posts categorized by different information sources, Facebook, Twitter and Pantip along with the timeline. The right panel of Figure 4 shows the overall summary of sentiment analysis in percentages of positive and negative polarities. Two tag

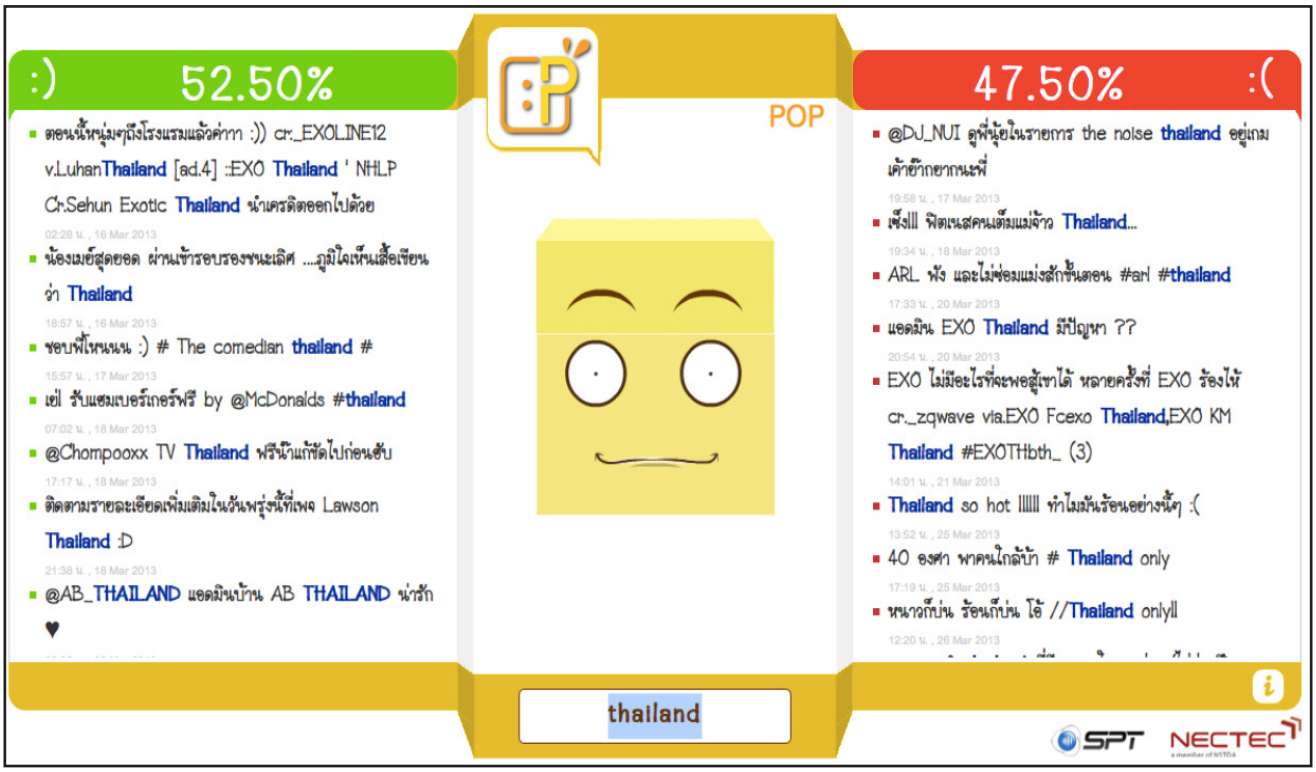

Figure 4. *POP: real-time social media monitoring application.*

clouds, feature and polar, display the most frequent extracted feature and polar terms, respectively. For example, the most frequent feature terms are “promotion”, “network”, “3G”, “SMS” and “sims”. The most frequent polar terms are “overpriced”, “free”, “having problem” and “changing service”.

Figure 4 shows a captured screen of a real-time social media monitoring application called POP. The POP application collects real-time Twitter feeds written in Thai language. The goal is to analyze the overall real-time sentiment of Twitter users in Thailand. User may also enter a keyword to search and filter for Twitter posts containing the specified keyword. Each Twitter post is sent to S-Sense engine to analyze the user’s sentiment. The percentages of positive and negative sentiments along with the analyzed posts are displayed on the left and and right panels, respectively. The animated and adaptive POP emoticon display 5 levels of sentiment from the most satisfied (happy) to the most unsatisfied (unhappy).

7. Conclusion

We proposed a framework called S-Sense (Social Media Sensing) for developing a social media analyzing tool. In the preprocessing step, we first proposed an algorithm for tokenizing and normalizing Thai written texts. The proposed algorithm DCB-Norm is a dictionary-based parser with a rule-based extension to merge and remove repeated characters. We evaluated the proposed algorithm by using a corpus of 1,000 manually tokenized texts. The accuracy is equal to 96.3% with the average throughput of 435,596 words/second.

For the intention and sentiment analysis, we applied the Naive Bayes as the classification algorithm to analyze four different intentions (announcement, request, question and sentiment) and two sentiments (positive and negative). The proposed framework was evaluated by using a social media corpus in the domain of mobile service obtained from Twitter and Pantip web board. To study the effect of using different lexicon sets to train the models, we compared two approaches: using only general lexicon and using both general lexicon and clue terms. The results showed that adding clue terms into feature vector for training the classification models helps improve the accuracy for all intention and sentiment analysis models. For intention models of request, question and sentiment, the accuracy is increased by approximately 6%. For sentiment model, the accuracy is equal to 91.64% an increase of approximately 2%. From the error analysis, we found that two major problems are word sense ambiguity and sarcasm.

S-Sense has been applied in many applications including social media monitoring and analytical in mobile service and tourism domains. Another interesting application area is employee engagement monitoring by analyzing comments of employees in enterprise social network platform. For useful real-time disaster management, S-Sense was applied as a tool to monitor and analyze the Twitter post streams during Thailand’s flooding crisis in the year 2010.

8. References

[1] X. Ding, Bing Liu, and P. S. Yu. “A holistic lexicon-based approach to opinion mining.” *In Proceedings of International Conference on web search and web data mining* , pp. 231-240, 2008.

[2] W. Jin, H. H. Ho, and R. K. Srihari. “OpinionMiner: a novel machine learning system for web opinion mining and extraction. ” *In Proceedings of the 15th ACM SIG KDD*, pp. 1195-1204, 2009.

[3] M. Tsytsarau and T. Palpanas. “Survey on mining subjective data on the web.” Data Minining and Knowledge Discovery, Vol. 24, No. 3, pp. 478-514, 2012.

[4] B. Pang, L. Lee and S. Vaithyanathan. “Thumbs up?: sentiment classification using machine learning techniques.” *In Proceedings of the ACL-02 Conference on empirical methods in natural language processing*, pp. 79-86, 2002.

[5] P. Beineke, T. Hastie, and S. Vaithyanathan. “The sentimental factor: improving review classification via human-provided information.” *In Proceedings of the 42nd Annual Meeting on Association for*

Computational Linguistics , pp. 263-270, 2004.

[6] S. M. Kim and E. Hovy. "Determining the sentiment of opinions." *In Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367-1373, 2004.

[7] T. Wilson, J. Wiebe, and P. Hoffmann. "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis." *Computational Linguistics*, Vol. 35, No. 3, pp. 399-433, 2009.

[8] M. Hu and B. Liu. "Mining and summarizing customer reviews." *In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pp. 168-177, 2004.

[9] L. W. Ku and H. H. Chen. "Mining opinions from the Web: Beyond relevance retrieval." *Journal of American Society for Information Science and Technology*, Vol. 58, No. 12, pp. 1838-1850, 2007.

[10] L. W. Ku, T. H. Huang, and H. H. Chen. "Using morphological and syntactic structures for Chinese opinion analysis." *In Proceedings of the 2009 Empirical Methods in Natural Language Processing*, pp. 1260-1269, 2009.

[11] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and C. Sangkeettrakarn. "Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews." *In Proceedings of the Eighth Workshop on Asian Language Resources*, pp. 64-71, 2010.

[12] A. Aw, M. Zhang, J. Xiao, and J. Su. "A phrase-based statistical model for SMS text normalization." *In Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pp. 33-40, 2006.

[13] M. R. Costa-Jussa and R. E. Banchs. "Automatic normalization of short texts by combining statistical and rule-based techniques." *Language Resources and Evaluation*, Vol. 47, No. 1, pp. 179-193, 2013.

[14] B. Han and T. Baldwin. "Lexical normalisation of short text messages: makn sens a #twitter." *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 368-378, 2011.

[15] B. Han, P. Cook, and T. Baldwin. "Lexical normalization for social media text." *ACM Transactions on Intelligent Systems and Technology*, Vol. 4, No. 1, pp. 1-27, 2013.

[16] F. Liu, F. Weng, and X. Jiang. "A broad-coverage normalization system for social media language." *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 1035-1044, 2012.

[17] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words." *Computer Speech and Language*, Vol. 15, No. 3, pp. 287-333, 2001.

[18] C. Asaga, Y. Mukarramah, and C.Watanabe. "ONOMATOPEDIA: onomatopoeia online example dictionary system extracted from data on the web. " *In Proceedingsof the 10th Asia-Pacific web Conference on Progress in WWW research and development*, pp. 601-612, 2008.

[19] A. Kato, Y. Fukazawa, T. Sato and T. Mori. "Extraction of onomatopoeia used for foods from food reviews and its application to restaurant search." *In Proceedings of the 21st International Conference oncompanion on World Wide Web*, pp. 719-728, 2012.

[20] Y. Uchida, K. Araki, and J. Yoneyama. "Classification of Emotional Onomatopoeias Based on Questionnaire Surveys." *In Proceedings of the 2012 International Conference on Asian Language Processing*, pp. 1-4, 2012.

[21] C. Haruechaiyasak and A. Kongthon. "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool." *In Proceedings of the 4th Workshop on South and Southeast Asian NLP (WSSANLP)*, pp. 9-16 2013.

[22] A. McCallum and K. Nigam. "A Comparison of Event Models for Naive Bayes Text Classification." *In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pp. 41-48, 1998.

[23] R. G. Ibáñez, S. Muresan, and N.Wacholder. "Identifying sarcasm in Twitter: a closer look."

In Proceedings of the 49th ACL: Human Language Technologies, pp. 581-586, 2011.

[24] F. Kunneman, C. Liebrecht, M. V. Mulken, and A. V. den Bosch, "Signaling sarcasm." *Information Processing and Management*, Vol. 51, No. 4, pp. 500-509, 2015.

[25] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm Detection on Twitter: A Behavioral Modeling Approach." *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, pp. 97-106, 2015.