Opinion Detection in Thai Political News Columns Based on Subjectivity Analysis

Khampol Sukhum*, Supot Nitsuwat**, and Choochart Haruechaiyasak***

Abstract

News columns are opinionated contents in newspapers such as editorials where columnists usually express their opinion and taking side explicitly with issues in the society. While this is normal in democratic society, in some places where journalism is often lack of, this bias could encourage rifts and polarity, especially in delicate issues such as politics. With such concern in mind, in this paper, we propose an opinion mining framework for detecting opinions in Thai political news columns. Our goal is to construct an information filtering model for monitoring highly-opinionated news content. Findings provided in this paper included our primarily result and discussion about difficulties found in applying opinion mining in Thai, as well as our analysis. In our work, contents from Thai news columns, annotated at sentence level by human experts, are used as training dataset to build a model in which various classifiers and feature selection techniques have been experimented. The best result is achieved by Multinomial Naïve Bayes with prior-knowledge based feature selection. The Precision, Recall, and F-Measure are equal to 0.809, 0.804, and 0.803 respectively.

Keywords: Opinion mining, Sentiment analysis, Media monitoring.

1. Introduction

Opinion mining is a field of research which emphasizes on detecting expressions, emotions, viewpoints and private states, expressed in contents [1], [2]. Because of its potentials, applications of opinion mining are many. Such examples are, a company could track customers tastes about their products, average consumer searching for product information could determine beforehand if a product is really worth buying, government agency can estimate its public policy easier by surveying expressions from online communities and social network websites, etc.

In recent years, opinion mining has been successfully applied in various domains, largely in reviews such as [3]. However, in news domain, the task is quite complicated compared to those in reviews. A study by [4] stated that news data is very much different from reviews: expressions are more subtle, features could be infinite, news pieces such as news quotation is often short, contains multiple targets, and has more varieties of affective expressions.

While opinion mining in English has been well-studied, in Thai, this field is still young. Thai differs from English greatly: it has no word and sentence boundary, it is anaphora language in which elements in a sentence or phrase can be drop without change in the meaning [5]. Thai Named Entities [6], are also common in news type contents. Together, these obstacles could make opinion mining in Thai become an interesting field.

Through our work, we have been trying to explore both issues: ambiguities in news contents and Thai language. This paper summarized our work in applying opinion mining for building opinion identification framework for Thai political news columns.

Our decision on selecting political news columns is based on many reasons. Firstly, similar to editorials, a political news column is usually a content of mixed facts and opinion. Columnists often make contempt or favorite remarks against entities; this makes them useful to help understand the characteristics of media bias phenomena. Secondly, at document level, a news column is likely to be categorized as opinion piece; so it is interesting to see if one could pinpoint exact location of opinion pieces in these news columns. Our aim is to do this at a sentence level. This is not new in English where sentence and word boundaries are less ambiguous; in Thai, it is a challenging task.

The goal of our proposed framework is to build information filtering system in which we can use subjective classifiers to identify and labeling opinion pieces in Thai political news

* *Department of Information Technology, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok.*

** *Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok.*

*** *Human Language Technology Laboratory (HLT) National Electronics and Computer Technology Center.*

columns. Our aim is to train such classifiers, which we emphasize on opinion identification, for the automatic labeling task at sentence level. We believe that, like labeling media with explicit content, once these opinion pieces were recognized, media consumers will be more aware to the nature of bias in the content. After all, an information filtering system based on this framework could be useful in a society where media neutrality and polarization is in concern.

In section 2 of this paper, we will discuss about background and related works in this field before we move on to section 3 where our methodology will be introduced, and then our experiments will be discussed in section 4. Finally, in section 5, we will conclude our findings and provide suggestion for future works.

2. Background

Web 2.0 has led to the emerging of opinion mining where main motivation is to let computer recognize emotions, expressions, sentiment, or private state in text [1], [2]. In order to achieve such goal, the Subjectivity Analysis approach as pioneered by [2] stated that relying on standard dictionary alone is insufficient, in order to recognize expressions, subjectivity clues have to be available. In their work, Potential Subjective Elements such as unique words, collocations, and POS, generated from manual annotation tasks, as described in detail in [7], are used in various combinations for recognizing subjectivity. The major drawback of the manual annotation approach is that it takes much time and efforts, thus many researchers decided to go for automatic approach instead. For example, [8] use a bootstrapping process to extract subjective sentences for subjective pattern extraction task. Extracted pattern is feed back to the bootstrapping process and thus repeat the system. [9] use similarity in contents for opinionated document recognition, sentence extraction and use co-occurrence of words with seed words for polarity identification task. For feature selection, while [10] focused on using topic-independent features such as writing style in newspaper for objectivity classification, [11] use hotel reviews as case study for syntactic pattern features extraction in building opinion-mining resources for Thai Language.

In political, newspaper, and quotation domain, the task is more difficult than other domain such as movie reviews, largely due to ambiguities found in content. [12] stated that mixed speeches, implicit opinion expressions, implicit targets, selective, partial target, and procedural speeches are difficulties found in political opinion expressions. [4] stated that degree of expressions, content length, subjectivity with ambiguity in targets are among obstacles in newspaper quotation polarity detection and argued that interpretation about polarity could be more than just negative versus positive viewpoints. The accuracy, based on F-Measure score, in both mentioned works are between 70 – 80%.

3. Method

A. Proposed Framework

Our proposed framework can be divided into three major parts: data collection, annotation, and classification.

At the data collection part, a document parser is responsible for collecting, parsing, cleaning, and storing raw data from online sources. A document prepared this way is ready to be used as a dataset for annotation task, or as a document to be identified for subjective pieces.

For building a corpus, at the second part, human experts are responsible for annotation tasks in which the purpose is to identify subjective elements as well as other useful features in the dataset taken from the first part.

At the last part - classification, a model is then build using a training dataset generated from a corpus based on human annotation at the second part. Later on, this model could be used for subjective pieces identification in a document.

The diagram of our framework is as shown in figure 1.

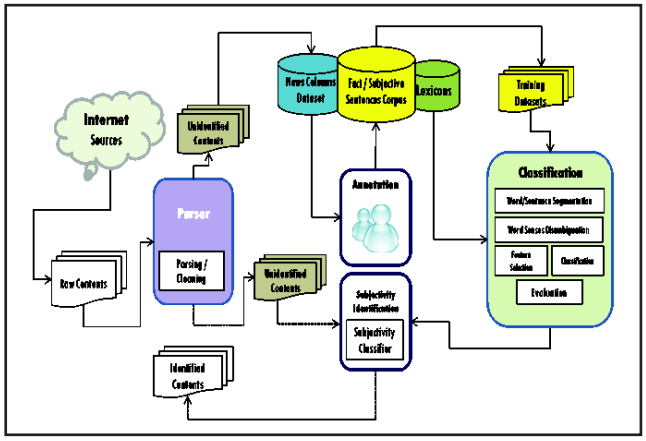

Figure 1 *Opinion Identification Framework for Thai political news columns.*

In next section we will describe our framework in detail

4. Experiment And Discussion

A. Data Collection

Our dataset is a collection of 117 pieces of political news columns written by 8 different columnists and published in 8 well-known Thai online newspapers: Thairath, Matichon, Daily News, Bangkok Business, Siamrath, Khaosod,

Komchadluek, and Manager Online, during the time period of June 2009 - September 2010. They are collected in approximately equal amount. Topics covered in these columns are mostly political, economic, social news and events; though there are some off-topic contents such as readings, and tourism as well. Unrelated elements, for examples, images, html tags, writer's name, newspaper title, etc. are removed; only content and title of each column are kept.

B. Annotation

We asked 5 Thai natives with language-related academic background to become our annotators. One of the annotators was asked to do sentence segmentation. By her effort, original 117 documents have been broken down into a total number of 2,539 sentences, with total vocabulary size estimated at around 62,000 words. This sentences-tagged dataset is categorized into 8 groups based on their sources, put into a single collection and given to all annotators. Finally, we asked our annotators to undertake marking of sentences that seemed to be subjective, according to their own judgment. In some clear-cut cases, we also asked them to undertake marking of subjective words in the subjective sentence.

Table 1 *samples of sentences with clue words marked.*

Sentence	Type
ความไม่ชอบมาพากลใน กกต.นั้นมีหลายเรื่อง เช่น หมกเม็ดการสอบสวนวินัยร้ายแรงเจ้าหน้าที่ กกต.2 คน ที่ปลอมลายเซ็นเอกสารที่ส่งศาลฎีกาในคดีใบแดงนายยงยุทธ ติยะไพรัช นานกว่าครึ่งปี แต่ขอยกมาเพียง 2 ตัวอย่าง ดังนี้	Subjective
เพราะฝ่ายตรงข้ามนั้นรอโอกาสอย่างยิ่งยวด หลังจากล้มไม่เป็นท่าทั้งจากการเมืองในรัฐสภาและการเมืองนอกระบบมาแล้ว	Subjective
นายประสาร ไตรรัตน์วรกุล เริ่มงานที่ ธปท.ตั้งแต่ปี 2526-2535 ในฝ่ายตรวจสอบและกำกับสถาบันการเงิน	Objective
เมื่อต้นสัปดาห์ที่ผ่านมา คณะกรรมาธิการติดตามสถานการณ์บ้านเมือง วุฒิสภา เชิญผู้บังคับการตำรวจใน 4 จังหวัดภาคตะวันออกเฉียงเหนือเข้าให้การ	Objective

C. Baseline building

1) Gold Standard Dataset

Sentences are selected based on majority votes in each objective/subjective category from a corpus previously built at annotation phase. Out of 2,539 original sentences in the corpus, a total amount of 1,380 sentences are selected for gold standard dataset. The dataset contains 30,952 words, with an average 22.429 words per sentence. 1,344 clue words are also collected.

2) Classification

For word segmentation task and POS tagging, we use LexToPlus1, a Thai word segmentation utility, and a 32,000 words size, LEXiTRON dictionary2, by National Electronics and Computer Technology Center (NECTEC), Thailand. To improve segmentation accuracy, collected subjective clues and named entities are supplied as parameters. We use Weka3 as our testing platform for all of the classification tasks. For the best classifier, we compared results from 3 well-known classifiers: SVM-SMO, Naïve Bayes Multinomial, and K-NN, to establish our baseline. Here, Naïve Bayes Multinomial is selected over standard Naïve Bayes for many reasons which will be discussed in next section. All classifiers are trained based on 10-fold cross validation, with default parameters. For feature selection, unigrams with TF weighting scheme is used. Weight of a feature is calculated by simply counting their appearances in a sentence, nothing is removed except numbers and symbols.

We achieved the best result with Naïve Bayes Multinomial and F-measure score equal to 0.793, for the baseline.

Table 2 *Baseline result: precion, recall, f-measure.*

Classifier	P	R	F
Naïve Bayes	0.795	0.793	0.793
SVM-SMO	0.75	0.749	0.749
K-NN	0.571	0.536	0.469

D. Improving Performance

1) Classifier

Since Naïve Bayes classifier has given the best performance in our baseline, it is selected as a classifier for improving further performance.

According to [13], there are two types of Naïve Bayes classifiers which have been used in machine learning: the standard model which is known as the multi-variate Bernoulli model, and the Multinomial model, or "the unigram language model". The standard model is based on assumption that there is no connection between words occurrences. Based on this model, the probability of a word is calculated based on its appearance and disappearance while number of time the word occurred in a document is discarded.

Unlike standard model, the multinomial model captures occurrences of words in a document. The probability of a word is multiplied by its occurrences in a document. The occurrences are referred as events in a document which is referred as collection of events. This model has been known

to give better performances over the standard model.

2) Features Selection

POS, Symbols, and Negation

Some early works suggested that POS such as adjectives or adverbs [2] are good subjective indicator. Other elements such as negation word, and Quotation symbols (‘ ” ’) have also been used. In our cases, adding weight to presence of adjectives, adverbs, or both together increased result just slightly. And there are sentences which contain not just quotation but also question (?), and exclamation (!) marks. The accuracy reduced slightly when adding weight to presences of Quotation mark; however, this trend is reversed with question and exclamation marks.

Stopwords

Stopwords are common words which can be frequently found in documents but are not useful for being used as features; traditionally, in text categorization tasks, these words are often discarded.

Table 3 shows Thai stopwords in our experiment.

Table 3 *Thai Stopwords.*

Thai Stopwords
ไว้,ไม่,ไป,ได้,ให้,ใน,โดย,แห่ง,แล้ว,และ,แรก,แบบ,แต่,เอง,เห็น,เลย,เริ่ม,เรา,เมื่อ,เพื่อ,เพราะ,เป็นการ,เป็น,เปิดเผย,เปิด,เนื่องจาก,เดียวกัน,เดียว,เช่น,เฉพาะ,เคย,เข้า,เขา,อีก,อาจ,อะไร,ออก,อย่าง,อยู่,อยาก,หาก,หลาย,หลังจาก,หลัง,หรือ,หนึ่ง,ส่วน,ส่ง,สุด,สำหรับ,ว่า,วัน,ลง,ร่วม,ราย,รับ,ระหว่าง,รวม,ยัง,มี,มาก,มา,พร้อม,พบ,ผ่าน,ผล,บาง,น่า,นี้,นำ,นั้น,นัก,นอกจาก,ทุก,ที่สุด,ที่,ทำให้,ทำ,ทาง,ทั้งนี้,ทั้ง,ถ้า,ถูก,ถึง,ต้อง,ต่างๆ,ต่าง,ต่อ,ตาม,ตั้งแต่,ตั้ง,ด้าน,ด้วย,ดัง,ซึ่ง,ช่วง,จึง,จาก,จัด,จะ,คือ,ความ,ครั้ง,คง,ขึ้น,ของ,ขอ,ขณะ,ก่อน,ก็,การ,กับ,กัน,กว่า,กล่าว

In our work, adding weight to their appearances slightly degraded the result (-0.4%) though not as much as removing them out (-0.7%).

We believed that stopwords do have their function in completed the semantic structure of a sentence, in either facts and opinion writing, their absence or presence could have impact in which how those sentence have to be written, but simply counting their appearance alone seemed not to be useful.

Numbers and Numerical elements

In many cases, factual information such as numbers, dates, events, people, and places, was referred in a sentences as supporting evidences by columnists, for examples:

จากข้อเท็จจริงดังกล่าวแสดงว่า กกต. อย่างน้อย 3 คน เห็นว่า นายบุญจงมีความผิดมากว่าการปราศรัยใส่ร้ายฯซึ่งน่าจะเพียงพอในการให้ใบแดงแล้ว.

Based on those mention facts, at least 3 members of the Election Committee agreed that Mr. Boonchong is guilty in cases severe than that of average defaming case, which definitely make him worth a red card by the committee.

เหตุระเบิดครั้งที่ 2 ที่หน้าห้างคิงเพาเวอร์ ดิวตี้ ฟรี คงไม่สามารถตามจับตัวผู้ก่อเหตุป่วนเมืองได้อย่างรวดเร็วเหมือนเช่น 2 ครั้งที่ผ่านมา.

[The perpetrator] of the 2nd bombing at King Power Duty Free shop will not be easily solved, unlike the last 2 incidences.

and,

ซึ่งเมื่อปี 2552 เคยเกิดขึ้น ต่อมาในปี 2553 ประวัติศาสตร์ก็ซ้ำรอยแถมหนักกว่าเดิม..That happened before in 2008 and again in 2009, which is much worse.

Numbers are discarded in previous study for they are not usually indicated subjectivity by nature [2]. In our cases, however, when more weight is added, their appearances improved the result.

Prior-Knowledge based features: Keywords, Named Entities, and Clue words

Looking closely at the corpus, we ‘ve found that in some sentences, columnists often stressed their point by wrapping quotation marks around keyword:

ผมคิดว่าตรงนี้คือ “จุดอ่อน” ในระบบการศึกษาของเรา. I Think, this is a “weak spot” in our education system.

Sometimes, words quoted like this implied writer’s opinion and we identified and extracted 397 pieces of them from our corpus. When more weight is added for their presences in a sentence like quotation and exclamation marks, these keywords improved accuracy as well.

Named Entities (NE) are common in our dataset. We found that 72% of sentences in our training dataset contain at least one NE, example:

ล่าสุดขุนคลังโชว์วิสัยทัศน์ให้สัมภาษณ์สื่อนิตยสารที่ทรงอิทธิพลของโลกอย่าง newsweek เกี่ยวกับแนวทางในการแก้ไขปัญหาวิกฤตการเมืองไทย ซึ่งนายกรณ์ จาติกวณิช นำเสนอออกมาสามแนวทางคือ Recently,the minister did a show-off of his vision, by giving interviews to the world’s powerful magazine – newsweek about his idea on resolving Thai political crisis in which he, the financial minister-Korn Jatikavanich, proposed the three exit strategies.

In our cases, removing NEs out of sentences degraded the result; this might be caused by drops in numbers of features in a sentences. We believed that the classifier relies on them for a clue in short sentences, similar to phenomena where objectivity classifiers trained on bag-of-words inclining on

learning topics as observed by [10]. While our human tagged clue words seems to be not clear-cut or frequent enough, still, adding weight to their presence increased result above that of NEs.

While grouping NEs together as a single feature - to reduce their effects on a sentence - resulted in slightly lower accuracy as oppose to grouping clue words, our best performance concerning features came from grouping both NEs and Clues, and feature weight is given based on their appearances.

3) Result

Table 4 *results from different features selection.*

	Features	P	R	F
Prior-Knowledge	Clues	0.802	0.799	0.799
	NEs	0.75	0.79	0.79
	Grouping Clues	0.814	0.802	0.8
	Grouping Clues, NEs	0.809	0.804	0.803
	Keywords	0.796	0.794	0.794
POS	Adj	0.798	0.796	0.796
	Adv	0.794	0.794	0.794
	Adj+adv	0.798	0.796	0.795
	Neg (ไม่,อย่า,มิ)	0.797	0.795	0.795
	Symbols '?', '!'	0.796	0.794	0.794
	Numbers	0.798	0.796	0.795

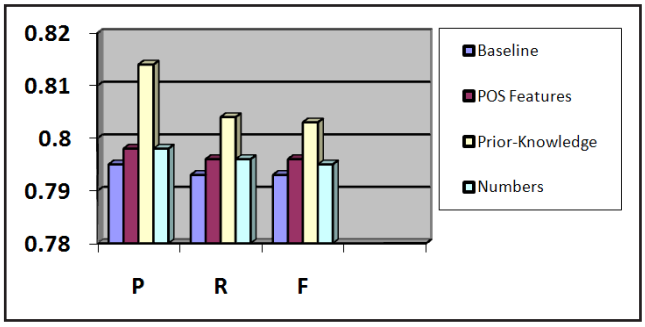

Figure 2 *Performance of features against baseline.*

E. Discussion

Our dataset contains both clear-cut subjective cases:

นี่คือปัญหาเรื่องการเรียน(เพียง)เพื่อสอบเข้ามหาวิทยาลัย และ วิธีการสอบเอ็นทรานซ์ที่งี่เง่า ที่ทำกันมาเป็นสิบปีแล้ว. This is a problem of educations just for entering college, and stupid entrance examination system which has been in place for a decade.

And the borderline cases, as stated in [2]:

เพราะนี่คือความจริงที่มีการพิสูจน์ให้เห็น ผ่านทางคำพิพากษาของศาล ซึ่งเป็นศาลยุติธรรมตามหลักสากล. Because this is a facts which are supported by verdicts from justice court which is established according to the rule of international law.

Disagreement rates among annotators seemed to be high in latter cases, while sentences which contain subtle subjectivity are likely to be marked as facts:

วันเสาร์สบายๆ วันนี้ ผมมีหนังสือเล่มหนึ่งจะแนะนำกับท่านผู้อ่าน. In this one fine Saturday, I have a book to introduce to my readers.

Our classification's performance suffered heavily from such borderline cases. Due to this ambiguity, features such as clue words alone seemed to be less effective.

Keywords improved precision in some sentences. But their extraction task can't be straightforward, since Quotation marks are also used for other purposes and manual extraction tasks are always expensive. Symbols such as question marks and exclamation marks have also helped increased accuracy, although we believed that such writing style might be a rare case.

Clue words and Named Entities sometime signaled the use of subjectivity. But their usefulness seemed to be ineffective due to infrequencies and ambiguity.

5. Conclusion

Our work has shown that, while basic features such as unigrams are effective enough for opinion identification task in news columns, their precision can be improved further with various features such as POS, special symbols, and prior-knowledge based features such as Clue words, Keywords, and Named Entities. Stopwords also have their usefulness in identifying opinionated sentences, though we believed that performing feature extraction by syntactical approach could be more suitable for these words.

Borderline cases also have effects on performance; this problem has to be addressed in further research. The most influence factors, in our opinion, are still features selection. While n-grams based features are good enough, domain-independent features [11] such as syntactic structures might improve further classification result especially in short sentences. This entire task requires deeper understanding in natural language processing.

Unfortunately, due to the cost of manual annotation tasks, our training dataset is quite small; we believe that, because of this, many useful features are not yet effective.

For our future work, we will investigate further into this direction and domain-independent features selection approach, as well as automatic corpus building for a larger dataset and larger feature base.

6. References

[1] B. Pang, L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, Vol.2 No.1-2, pp. 1-135, January, 2008.

[2] J. Wiebe, T. Wilson, R. Bruce, M. Bell, M. Martin, "Learning Subjective Language," *Computational Linguistics*, Vol.30 No.3, pp. 277-308, 2004.

[3] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal taxonomies for sentiment analysis," *Proceedings of the 2nd Midwest Computational Linguistic Colloquium*, Columbus, US, 2005.

[4] B. Alexandra, S. Ralf, v. d. Goot Erik, P. Bruno, K. Mijail, "Opinion Mining on Newspaper Quotations," *Proceeding of Int. Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2009.

[5] C. Jirawan, K. Asanee, "Thai Elementary Discourse Unit Segmentation by Using Discourse Segmentation Cues and Syntactic Information," *Master of Engineering Thesis*, Computer Engineering Dept. Kasetsart University, 2006

[6] T. Nattapong, T. Thanaruk, "Pattern-based Extraction of Named Entities in Thai News Documents," Thammasat Int. J. Sc. Tech., Vol.15 No.1, 2010.

[7] J. Wiebe, T. Wilson, and C Cardie, "Annotating expressions of opinions and emotions in language," Language Resources and Evaluation, Vol. 39 No. 2-3, 2005.

[8] E. Riloff , J. Wiebe, "Learning extraction patterns for subjective expressions," *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105-112, 11 July, 2003.

[9] H. Yu , V. Hatzivassiloglou, "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences", *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 129-136, 11 July, 2003.

[10] E. Lex, A. Juffinger, M. Granitzer, "Objectivity classification in online media," *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 13-16 June, Toronto, Ontario, Canada, 2010.

[11] C. Haruechaiyasak, A. Kongthon, P. Palingoon and C. Sangkeettrakarn, "Constructing Thai Opinion Mining Resource: A Case Study on Hotel Reviews," *Proceedings of the 8th Workshop on Asian Language Resouces*, pp. 64-71, Beijing, China, 2010.

[12] B. Yu , S. Kaufmann , D. Diermeier, "Exploring the characteristics of opinion expressions for political opinion classification," *Proceedings of the* 2008 *international conference on Digital government research*, Montreal, Canada, 18-21 May, 2008.

[13] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classificatio," *In AAAI*-98 *Workshop on Learning for Text Categorization*, 1998.