Emotion Classification of Thai Text based Using Term weighting and Machine Learning Techniques

Nivet Chirawichitchai
Faculty of Information Technology, Sripatum University,Thailand.
nivet99@hotmail.com

***Abstract*—In this research, I proposed Emotion Classification of Thai Text based Using Term weighting and Machine Learning Techniques focusing on the comparison of various common term weighting schemes. I found Boolean weighting with Support Vector Machine is most effective in our experiments. I also discovered that the Boolean weighting is suitable for combination with the Information gain feature selection method. The Boolean weighting with Support Vector Machine algorithm yielded the best performance with the accuracy over all algorithms. Based on our experiments, the Support Vector Machine algorithm with the Information gain feature selection yielded the best performance with the accuracy of 77.86%. Our experimental results also reveal that feature weighting methods have a positive effect on the Thai Emotion Classification Framework.**

Keywords*— *Emotion Classification, Feature Reduction, Machine Learning.

I. Introduction

With the popularity of the Internet, the amount of comments in text based communication mechanisms such as webblogs, twitter, facebook, webboard and so on is going to dramatically increase every day; moreover, there has been growing interest in studying the methods through which emotion is expressed in text based communication. Whether it is mining for consumers' opinions, or tracing their intents of writers towards various topics from evaluations and inclinations at discovering sentiment, spotting and interpreting emotion from text based is highly applicable to various natural language processing areas. Some important examples include word sense disambiguation, multi-document summarization, and multi-perspective question answering. Previous work of this research in sentiment analysis has been done on a variety of text genres, including product and movie reviews news stories, editorials and opinion articles, and more recently, blogs, webboard positive or negative orientation; however, present work of this research focused on sentiment classification methods to capture emotion and opinion mining from text based of the natural language processing, specifically for only Thai language.[1, 2]

This paper describes sentiment classification experiments in the Emotion domain [3, 4] by creating a model that classified integrates the Term weighting with machine learning techniques. This study propose Emotion Classification of Thai Text based Using Term weighting and Machine Learning Techniques. The rest of the paper is organized as follows. Section 2 describes the feature extraction methods. Section 3 describes the Dimensionality Reduction. Section 4 describes the classification algorithms for empirical validation. Section 5 presents Thai Emotion Classification framework. Section 6 presents the experiments and results. Finally, Section 7 conclusions.

II. Feature Extraction

A. Dataset

The data set of this research is collected from various Thai popular social network webboard posts; for example, www.facebook.com, www.pantip.com, www.siamphone.com, www.kapook.com, www.sanook.com. Data set collected from the posting on these websites are classified into six basic emotional categories. Each emotional category represents the distinctly identifiable facial expressions of emotion – joy, sadness, anger, love , surprise and fear. [3, 4] Finding words commonly present the context of a particular emotion. Thus, these seed words could be categorized in many emotions, two samples of the emotions are joy ("ความสุขhappy", "สนุกenjoy" and "ยินดีpleased") and fear ("กลัวafraid", "ผวาscared", "ตื่นตกใจ panic"). However, we decided to focus our attention on webboard for two main reasons. First, webboard post have typically a high load of emotional content, as they describe major events. Second, the structure of webboard post was appropriate for our goal of conducting sentence-level annotations of emotions. The dataset consists of 1800 thai webboard posts, collected between november 1, 2013 and February 15, 2014.

B. Preprocessing

The first step in sentiment classification is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. For Thai language, the main task of text processing is the segmentation of texts into word tokens. Thai texts are naturally unsegmented, i.e.,words are written continuously without the use of word delimiters. Due to this distinct characteristic, preparing a feature set for Thai sentiment classification is more challenging than Latin based languages such as English, French and Spanish. With Latin-based languages, a text string can easily be tokenized into terms by observing the word delimiting characters such as spaces, semicolons, commas, quotes, and periods. To prepare a feature set for Thai documents corpus, we must first apply a word segmentation algorithm to tokenize text strings

978-1-4799-5822-1/14/$31.00 ©2014 IEEE

into series of terms. Once a set of extracted words are obtained from the training news corpus, the removal of HTML tags, removal of stop-words and then word stemming. The stop-words are frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). By word stemming we mean the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as walk, walker, walked, and walking. [5]

C. *Term weighting*

All Emotion documents are segmented into words or tokens that are inputs for next steps. In the vector space model, documents are represented by vectors of words. Term weighting method aims to indicate the significant of a term in a document. In Emotion classification, Boolean is simplest approach to let the weight be 1 if the word occurs in the document and 0 otherwise. TF and TF-IDF are widely applied to count the weight of a term. TF represents the number of times a term occurs in a document, and TF-IDF is the combining of TF and IDF weights. IDF indicates the general importance of a term in overall documents.[6] If a term's score of TF-IDF is high, it means this term occurs frequently and only appears in the part of overall documents. IDF and TF-IDF can be calculated as equations (1) and (2)

$$idf = \frac{the\,number\,of\,\,total\,documents}{the\,number\,of\,\,documents\,include\,a\,term} \tag{1}$$

$$TFIDF = tf * idf \tag{2}$$

III. DIMENSIONALITY REDUCTION

With increasing of the textual data in cyberspace, how to extract significant information from a huge amount of data have been become a serious problem. The objective of feature selection is to extract the important terms in the documents, and achieve the goal of dimension reduction. Emotion classification can be considered as a classification approach in machine learning in which features are words or terms extracted from a given text corpus. A problem in statistical Emotion classification is the high dimensionality of the feature space. There exists one dimension for each unique word found in the collection of documents, typically tenthousand. Hence, there is a need for a reduction of the original feature set, which is commonly known as dimensionality reduction in the pattern recognition literature. Most of the dimensionality reduction approaches can be classified into feature selection. Therefore, I applied feature selection technique by information gain. [7, 8] Information gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. The information gain of term t and category c_i is defined to be:

$$IG(t, c_i) = \sum_{c \in \{c_i, \overline{c}_i\}} \sum_{t' \in \{t, \overline{t}\}} P(t', c) \cdot log \frac{P(t', c)}{P(t') \cdot P(c)} \tag{3}$$

which are functions of the following four dependency tuples:

(t, c_i) : presence of t and membership in c_i.

$(t, \overline{c}_i)$: presence of t and non-membership in c_i.

$(\overline{t}, c_i)$: absence of t and membership in c_i.

$(\overline{t}, \overline{c}_i)$: absence of t and non-membership in c_i.

where: t and c_i represent a term and a category respectively. The first and last tuples represent the positive dependency between t and c_i, while the other two represent the negative dependency.

IV. CLASSIFICATION ALGORITHMS

After feature selection and transformation Thai emotional data sets can be easily represented in a form that can be used by a machine learning algorithm. Many text classifiers have been proposed in the literature using machine learning techniques, probabilistic models, etc. They often differ in the approach adopted: decision trees, naive-bayes, rule induction, neural networks, nearest neighbors, and lately, support vector machines. Although many approaches have been proposed, automated text classification is still a major area of research primarily because the effectiveness of current automated text classifiers is fault and still needs improvement. The goal of classification is to build a set of models that can correctly predict the class of the different objects. Once such a predictive model is built, it can be used to predict the class of the objects for which class information is not known. The key advantage of supervised machine learning methods over unsupervised machine learning methods is that by having an explicit knowledge of the classes the different objects belong to, these algorithms can perform an effective feature selection which leads to better prediction accuracy. This section provides a brief introduction to four well-known algorithms that are widely used for Emotion Classification i.e. Support Vector Machine, Naive Bayes, Decision Tree and K-Nearest Neighbor.

A. *Support Vector Machine (SVM)*

The idea of theSVM algorithm [9, 10] is based on the structure risk minimization principle .It has been shown in previous works to be effective for text classification. SVM divides the term space into hyperplanes or surface separating the positive and negative training samples. The SVM algorithm is to find the decision surface that maximizes the margin between the data points of the two classes. Following our results and previously published studies in text classification, we limit our discussion to linear SVM. The dual form of the linear SVM optimisation problem is to maximize :

$$\alpha^* = maximise_{\alpha} \sum_{i=1}^{1} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i x_j), \tag{4}$$

$$\text{Subject to } \sum_{i=1}^{l} y_i \alpha_i = 0, 0 \le \alpha_i \le C, i = 1 \ldots l$$

with α_i the weight of the examples and C the relative importance of the complexity of the model and the error.

B. *Naive Bayes (NB)*

The idea of the NB classier [11] is to use a probabilistic model of text. To make the estimation of the parameters of the model possible, rather strong assumptions are incorporated. In the following, word-based unigram models of text will be used, i.e. words are assumed to occur independently of the other words in the document. Let $P(y_i)$ be the prior probability of the class y_i and $P(a'_j \mid y_i)$ be the conditional probability to observe attribute value a'_j given the class y_i. Then, a naive Bayes classifier assign to a data point x' with attributes $(a'_1....a'_d)$ the class $\hat{\Phi}(x')$ maximizing :

$$\hat{\Phi}(x') = argmax_{y_i \in c} P(y_i) \prod_{j=1}^{d} P(a'_j \mid y_i) \quad (5)$$

C. Decision Tree (DT)

The idea of the DT algorithm [12] is a common method used in data mining. A DT classifier is a tree in which internal nodes are labeled by attributes (words occurrences in the case of text categorization), branches departing from them are labeled by tests the weight that attribute has in the test document, and leafs are labeled by categories. DT categorizes a test document by recursively testing the weights that the attributed labeling the internal nodes have in document vector, until a leaf reached. The most common approach to inducing a decision tree is to partition the labeled examples recursively until a stopping criterion is met. The partition is defined by selecting the test which divide all examples to the disjoint subsets assigned to the test branches, passing each example to the corresponding branch, and treating each block of the partition as a subproblem, for which a subtree is build recursively. A common stopping criterion for a subset of examples is that they all have the same class. Since the misclassification probability for the given example can be estimated according to the misclassification probability computed for the leaf that covers this example, we have directly computed confidence of the prediction as the Laplace estimation of the leaf's misclassification error:

$$conf_i = \frac{p+0.5}{p+n+1} \quad (6)$$

where p (n) is the number of positive (negative) training examples assigned to the leaf.

D. K-Nearest Neighbor (KNN)

The idea of the KNN algorithm [13] is a method for classifying objects based on closest training examples in the feature space. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. Given a test point, a predefined similarity metric is used to find the k most similar points from the train set. For each class y_i, we sum the similarity of the neighbors of the same class. Then, the class y_i with the highest score is assigned to the data point x' by the k nearest neighbors algorithm.

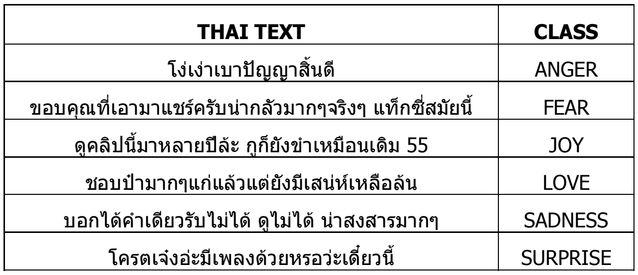
$$\hat{\Phi}(x') = argmax_{y_i \in c} \sum_{i=1}^{k} \delta(y_i, \Phi(x_i)) sim(x_i, x')$$

(7)

V. EMOTION CLASSIFICATION FRAMEWORK

In my research, I propose Emotion Classification of Thai Text based using Term weighting and Machine Learning Techniques framework, which consists of seven main steps:

1) The experiment using a collection of Thai Emotion Text obtained from the Thai popular social network webboard posts on Internet including www.facebook.com, www.pantip.com, www.siamphone.com, www.kapook.com, www.sanook.com. The natural language is simply used in the source of text, all emoticons and typographical symbols are ignored. A number of Emotion sentences are 1,800 records.



THAI TEXT	CLASS
โง่เง่าเบาปัญญาสิ้นดี	ANGER
ขอบคุณที่เอามาแชร์ครับน่ากลัวมากๆจริงๆ แท็กซี่สมัยนี้	FEAR
ดูคลิปนี้มาหลายปีล๊ะ กูก็ยังขำเหมือนเดิม 55	JOY
ชอบป๋ามากๆแก่แล้วแต่ยังมีเสน่ห์เหลือล้น	LOVE
บอกได้คำเดียวรับไม่ได้ ดูไม่ได้ น่าสงสารมากๆ	SADNESS
โครตเจ๋งอ่ะมีเพลงด้วยหรอว่ะเดี๋ยวนี้	SURPRISE

Fig. 1. Thai Emotion Dataset

2) Then 3 readers are asked to label the best emotion that can be exposed from the text. There are six emotions class: anger, fear, joy, love, sadness, surprise.
3) Next, pre-processed by the text processing. For Thai language, the main task of text processing is the segmentation of texts into word tokens, I must first apply a word segmentation algorithm to tokenize text strings into series of terms. I used a state-of-the-art word segmentation program called Kucut[14]which is based on unsupervised machine learningalgorithm as a tokenizer in this Bag-Of-Words approachand identify Part-Of-Speech Tagging.
4) Once a set of extracted words are obtained from the Thai Emotion Text based corpus, the removal of stop words and stemming from the dictionary begins.This step the words filtering is performed by selecting only nouns and verbs.
5) The output from this step will be used in the weighting scheme to assign the feature values as described in Section 2.
6) Reduce the number of word features by applying the feature selection technique as described in Section 3, by using the information gain statistics ranking for feature selection, the top p features per category were selected from the training sets.
7) For classifying emotion step, I used weka [15] an open-source machine learning tool, to perform the experiments. I used the default settings for all algorithms. For Support Vector Machine algorithm, the default kernel function is Linear kernel. The input comes from Emotion documents pre-classified into a set of Emotion class. Figure 4 illustrates the Thai Emotion Classification Framework.

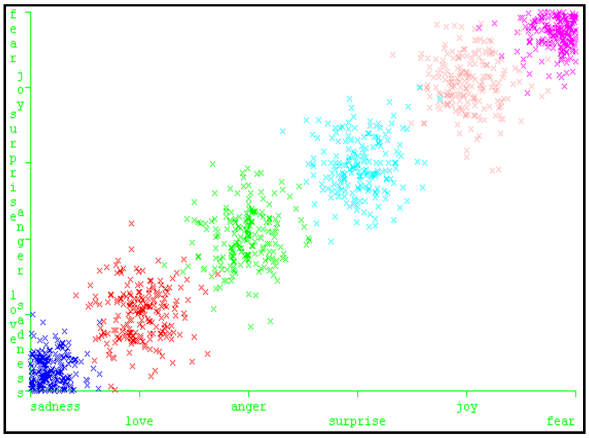


Fig. 2. Features in each class.

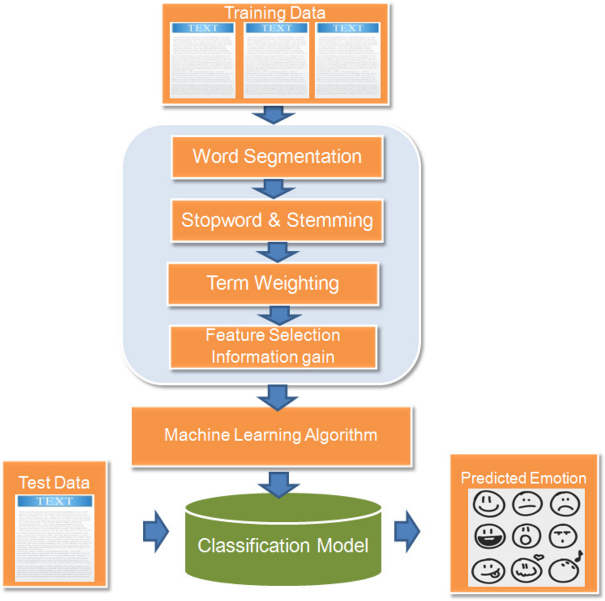
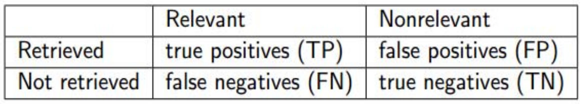


Fig. 3. Thai Emotion Classification Framework.

Classification effectiveness is usually measured using accuracy. Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. The accuracy is the proportion of true results (both true positives and true negatives) in the population. Accuracy is the fraction of decisions (relevant/non relevant) that are correct, function is computed as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (8)$$



	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

Fig. 4. Confusion Matrix

VI. EXPERIMENT AND RESULTS

I tested all algorithms using the 10-fold cross validation. The results in terms of accuracy are the averaged values calculated across all 10-fold cross validation experiments. The experimental results of these term weighting with respect to accuracy on Thai Emotion Text based dataset in combination with four learning algorithms are reported from Figure 5-8.

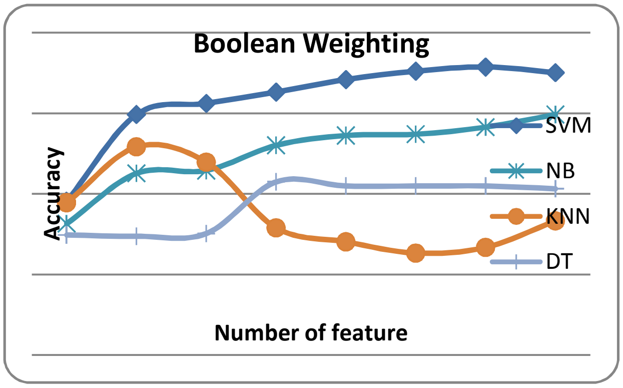


Fig. 5. The experimental results from Boolean weighting.

In figure 5 summarizes the results of classification with the Information gain feature selection method using four learning algorithms on Thai Emotion dataset after Boolean weighting. Five observations from the emotion classification of thai text using feature reduction and machine learning framework were found. First, Support Vector Machine algorithm is the most accurate, followed by subordinate Naïve-Bayes, Decision Tree, and K-Nearest Neighbor algorithms respectively. Second, Support Vector Machine, Naïve-Bayes and Decision Tree has a trend of the accuracy of classification increases as the number of the feature grows. Third, performance of the different learning algorithms with a small feature size can not be summarized in one sentence but the trends are distinctive that the accuracy points of different learning algorithms increase as the number of the features grows. Fourth, with the exception, K-Nearest Neighbor has an opposite direction of the trend of accuracy. That is, the accuracy significantly decline when feature increases. Finally, the best accuracy points of all algorithms is found in Support Vector Machine algorithm at 77.86% with feature size of 600.

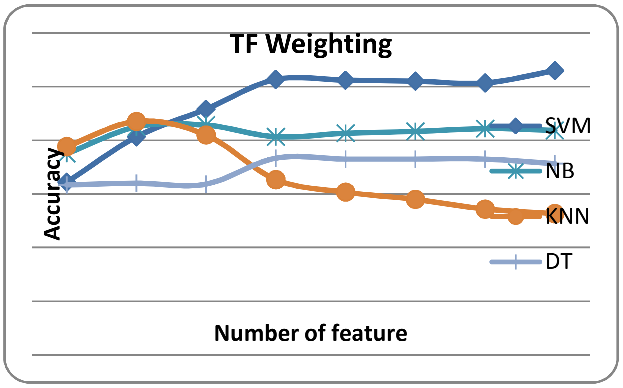


Fig. 6. The experimental results from TF weighting.

In figure 6 summarizes the results of classification with the Information gain using four learning algorithms on Thai Emotion dataset after TF weighting. Five observations from the Emotion Classification were found. First, Support Vector Machine algorithm is the most accurate, followed by subordinate Naïve-Bayes, Decision Tree, and K-Nearest Neighbor algorithms respectively. Second, Support Vector Machine, Naïve-Bayes and Decision Tree have a similar trend of stable accuracy. It shows that the reduction of the feature does not affect the accuracy of Emotion Classification. Third, Support Vector Machine has a trend of the accuracy of classification increases as the number of the feature grows. Fourth, with the exception, K-Nearest Neighbor has an opposite direction of the trend of accuracy. That is, the accuracy significantly decline when feature increases. Finally, the best accuracy points of all algorithms is found in Support Vector Machine algorithm at 76.50% with feature size of 700.

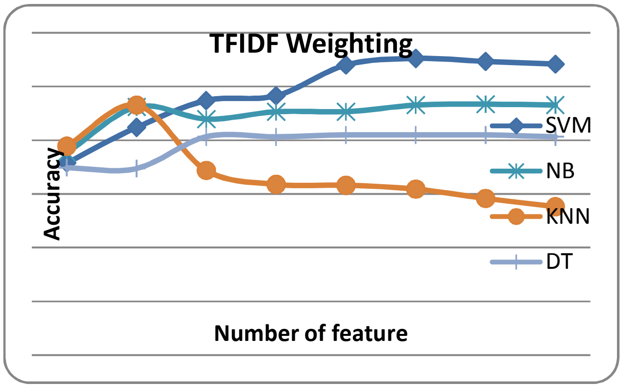


Fig. 7. The experimental results from TFIDF weighting.

In figure 7 summarizes the results of classification with the Information gain using four learning algorithms on Thai Emotion dataset after TFIDF weighting. Three observations from the Emotion Classification were found. First, Support Vector Machine algorithm is the most accurate, followed by subordinate Naïve-Bayes, Decision Tree, and K-Nearest Neighbor algorithms respectively. Second, performance of the different learning algorithms with a small feature size can not be summarized in one sentence but the trends are distinctive that the accuracy points of different learning algorithms increase as the number of the features grows. Finally, the best accuracy points of all algorithms is found in Support Vector Machine algorithm at 77.60% with feature size of 500.

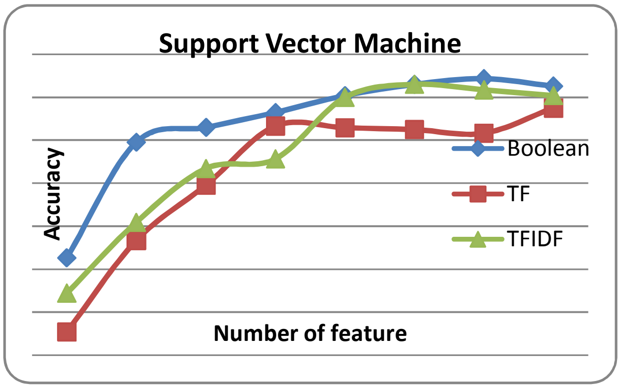


Fig. 8. The experimental results from Support Vector Machine.

In figure 8 summarizes the results of classification with the Information gain using Support Vector Machine algorithms on Thai Emotion dataset after Term weighting via Boolean, TF, TFIDF weighting, respectively. Four observations from the Emotion Classification were found. First, Boolean weighting is more effective than another weighting with Support Vector Machine on Thai Emotion Classification. Second, all term weighting schemes reached a maximum of accuracy point at the full feature. Third, the best accuracy points on Boolean weighting with Support Vector Machine were 77.86% at a feature size of 600. Finally, the Boolean weighting is suitable for Emotion Classification of Thai Text based more than the other weighting.

VII. Conclusions

In this research, I proposed Emotion Classification of Thai Text based Using Term weighting and Machine Learning Techniques focusing on the comparison of various common term weighting schemes. I found Boolean weighting with Support Vector Machine is most effective in our experiments. I also discovered that the Boolean weighting is suitable for combination with the Information gain feature selection method. The Boolean weighting with Support Vector Machine algorithm yielded the best performance with the accuracy over all algorithms. Based on our experiments, the Support Vector Machine algorithm with the Information gain feature selection yielded the best performance with the accuracy of 77.86%. Our experimental results also reveal that feature weighting methods have a positive effect on the Thai Emotion Classification Framework.

Acknowledgment

The authors would like to thank Sripatum University for scholarship support.

References

[1] Pang, B., Lee, L., and Vaithyanathan, S., “Thumbs up?: sentiment classification using machine learning techniques,” Proceedings of the conference on Empirical methods in natural language processing , Vol. 10, 2002 .

[2] Pang, B., and Lee, L., “Opinion Mining and Sentiment Analysis,” Found. Trends Inf. Retr, Vol. 2, pp. 1-135, 2008.

[3] Gill, A.J., French, R.M., Gergle, D., and Oberlander, J., “The language of emotion in short blog texts,” Proceedings of the ACM conference on Computer supported cooperative work, San Diego, CA, USA, 2008.

[4] Alm, C.O., Roth, D., and Sproat, R. , “Emotions from text: machine learning for text-based emotion prediction,” Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005.

[5] Haruechaiyasak, C., Jitkrittum, W., Sangkeettrakarn, C., and Damrongrat, C., “Implementing News Article Category Browsing Based on Text Categorization Technique,” Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 03, 2008.

[6] O’Keefe, T., and Koprinska, I., “Feature Selection andWeighting Methods in Sentiment Analysis,” Proceedings of 14th Australasian Document Computing Symposium, 2009.

[7] Nicholls, C., and Song, F., “Comparison of Feature Selection Methods for Sentiment Analysis,”Advances in Artificial Intelligence, Lecture Notes in Computer Science, Vol. 6085, pp. 286-289,2010.

[8] Sharma, A., and Dey, S., “A comparative study of feature selection and machine learning techniques for sentiment analysis,” Proceedings of the ACM Research in Applied Computation Symposium, San Antonio, Texas, 2012 .

[9] Burges, C.J.C., “A Tutorial on Support Vector Machines for Pattern Recognition,” Data Mining and Knowledge Discovery, Vol.2, pp. 121-167, 1998.

[10] Yang, Y., Xu, C., and Ren, G., “Sentiment Analysis of Text Using SVM,”Electrical, Information Engineering and Mechatronics 2011, Lecture Notes in Electrical Engineering, Vol. 138, pp. 1133-1139, 2012.

[11] Narayanan, V., Arora, I., and Bhatia, A., “Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model,”Intelligent

Data Engineering and Automated Learning, Lecture Notes in Computer Science, Vol. 8206, pp.194-201, 2013.

[12] Sui, H., Khoo, C., Chan, S., and Chan, S., "Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction,"14th Annual ASIST SIG CR Workshop, 2003.

[13] Tan, S., and Zhang, J., "An empirical study of sentiment analysis for chinese documents," Expert Systems with Applications,Vol.34, pp. 2622–2629, 2008.

[14] http://naist.cpe.ku.ac.th/pkg/kucut-1.2.2_python25_fix.zip

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., "The WEKA data mining software: an update," ACM SIGKDD Explorations Newsletter archive, Vol. 11, pp.10-18, 2009.

.