THAI SENTIMENT ANALYSIS ON SOCIAL MEDIA USING MAJORITY VOTING-BASED ENSEMBLE METHOD

Narin Panawas*

ABSTRACT

In this paper, we proposed Thai sentiment analysis on social media using majority voting-based ensemble classifier focusing on various term weighting schemes and multiple learning algorithms. We found majority voting-based ensemble method most effective in our experiments when comparison with single classifier such as naive Bayes, K-nearest neighbor and decision tree algorithm. We also discovered that the majority voting-based ensemble classifier is suitable for combination with the various term weighting on Thai sentiment analysis dataset. The majority voting-based ensemble method yielded the best performance with the accuracy over all traditional algorithms. Based on our experiments, the majority voting-based ensemble method with Boolean weighting yielded the best performance with the accuracy of 76.04%. Our experimental results also reveal that ensemble method have a positive effect on the Thai sentiment analysis based on social media framework.

Keywords: Thai sentiment analysis, ensemble classifier, term weighting.

INTRODUCTION

Recently in the area of machine learning the concept of combining classifiers is proposed as another course for the change of the direction for the improvement of the performance of individual classifiers. The idea of combining classifiers, in the area of machine learning, is set forth as a new way for improved performance of individual classifiers. These classifiers could be based on a variety of classification methodologies, and could achieve different rate of correctly classified individuals. The goal of classification result integration algorithms is to generate more certain, precise and accurate system results. Dietterich (Dietterich, 2000) provides an

* Lecturer, School of Information Technology, Sripatum University-Chonburi Campus

accessible and informal reasoning, from statistical, computational and representational viewpoints, of why ensembles can improve results.

The primary thought of ensemble methodology is to join a set of unique models, each of which solves the same original task, in order to obtain a better composite global model, with more accurate and reliable estimates or decisions than can be obtained from using a single model. Building a predictive model by coordinating various models has been under consideration for a very long time.

A group of researchers (Buhlmann & Yu, 2003; Dimitriadou, Weingessel & Hornik, 2003; Dietterich, 2000) pointed out that the history of ensemble methods starts as a long time ago, an ensemble methods can be also used for improving the quality and robustness of classification and clustering algorithms. The ensemble methods gives an available and reasoning, from statistical, computational and representational perspectives, of why ensembles can enhance results. In any case, in this part we concentrate on classifier ensemble methods.

In the previous few years, experimental studies conducted by the machine learning community show that combining the outputs of multiple classifiers reduces the generalization error,ensemble methods are very proficient, mainly due to the phenomenon that various types of classifiers have different inductive biases (Geman, Bienenstock & Doursat, 1995) Indeed, ensemble methods can effectively make use of such variety to reduce the variance error without increasing the bias error (Tumer & Ghosh, 1996). In some situations, an ensemble can also reduce error, as shown by the theory of large margin classifiers (Bartlett & Shawe-Taylor, 1998).

The ensemble methods is applicable in many fields such as: finance (Leigh, Purvis & Ragusa, 2002), bioinformatics (Tan, Gilbert & Deville, 2003), healthcare (Mangiameli, West & Rampal, 2004), geography (Bruzzone, Cossu & Vernazza, 2004), Sentiment analysis (Fersini, Messina & Pozzi, 2014), intrusion detection (Gaikwad & Thool, 2015) etc. Given the potential usefulness of ensemble methods, it is not surprising that a vast number of methods is now available to specialists, researchers and professionals.

This paper describes sentiment analysis experiments focusing in the emotional domain by creating a model that classified integrates the term weighting with ensemble classifiers methods. This study proposes the thai sentiment analysis on social media

using majority voting-based ensemble method. The rest of the paper is organized as follows. Section 2 describes the feature extraction methods. Section 3 describes the feature selection. Section 4 describes the Ensemble Classifiers Techniques for empirical validation. Section 5 presents thai sentiment analysis on social media using majority voting-based ensemble method framework. Section 6 presents the experiments and results. Finally, Section 7 conclusions.

FEATURE EXTRACTION

The data set of this research is collected from various Thai popular social network; for example, www.facebook.com, www.pantip.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 distinct identifiable facial expressions of emotion – joy, sadness, anger, love, surprise and fear. Finding words commonly present the context of a particular emotion. The dataset consists of 1800 Thai social network webboard posts, collected between January 1, 2016 and August 31, 2016.

1. Text Preprocessing

Data preprocessing is an important task and critical step in sentiment analysis, opinion mining, text mining, natural language processing and information retrieval. In the area of sentiment analysis, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data. The first step in sentiment analysis 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 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.). Stemming has the effect of mapping several morphological forms of words to a common feature. For example the words "learner", "learning", and "learned" would all map to the common stem "learn", and this latter string would be placed in the feature set rather than the former three (Polettini, 2004).

2. Term Weighting

The vector space model procedure can be divided in to three stages. The first stage is the document indexing where content bearing terms are extracted from the document text. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure. All sentiment documents in this research 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 sentiment analysis, 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 (Salton, Wong & Yang, 1975; Chirawichitchai, 2015). 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

$$\text{Idf} = \frac{\textit{the number of total documents}}{\textit{the number of documents include a term}} \quad (1)$$

$$\text{TFIDF} = \text{tf} * \text{Idf} \quad (2)$$

FEATURE SELECTION

This is a difficult question that may require deep knowledge to the problem sentiment analysis domain. It is possible to automatically select those features in your

data that are most useful or most relevant for the problem you are working on. This is a process called feature selection. A central problem in statistical sentiment analysis or text mining is the high dimensionality of the feature space, standard classification techniques cannot deal with such a large feature set, since processing is extremely costly in computational terms, and the results become unreliable due to the lack of sufficient training data. 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 (Yang & Pedersen, 1997). 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. It is frequently used as a term goodness criterion in machine learning. It measures the number of bits required for category prediction by knowing the presence or the absence of a term in the document. It is defined by following expression:

$$IG(t) = \begin{array}{l} -\sum_i Pr(c_i)\log Pr(c_i) + Pr(t)\sum_i Pr(c_i|t)\log Pr(c_i|t) \\ +Pr(\bar{t})\sum_i Pr(c_i|\bar{t})\log Pr(c_i|\bar{t}) \end{array} \qquad (3)$$

ENSEMBLE CLASSIFIERS TECHNIQUES

1. Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their conceivable consequences, including chance event outcomes, resource costs, and utility. It is one approach to show an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a scheme most likely to reach a goal, but are also a popular tool in machine learning. The core algorithm for building decision trees by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. Decision tree uses Entropy and Information Gain to construct a decision tree. Decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar homogenous,this algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one (Quinlan, 1987).

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (4)$$

And then after calculated entropy using the frequency table of attributes, the information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches). The information gain of attribute A, relative to a collection of examples, S, is calculated as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

2. Naive Bayes

The Naive Bayes (Rish, 2001) algorithm has been widely used for sentiment classification, and shown to produce very good performance. The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. Naive Bayes algorithm computes the posterior probability that the document belongs to different classes and assigns it to the class with the highest posterior probability. The naive part of Naive Bayes algorithm is the assumption of word independence that the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. The idea of the NB classier is 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 that is words are assumed to occur independently of the other words in the document. Let be the prior probability of the class and be the conditional probability to observe attribute value given the class . Then, a Naive Bayes classifier assign to a data point with attributes the class maximizing:

$$\hat{\Phi}(x') = argmax_{y_i \in c} P(y_i) \prod_{j=1}^{d} P(a'_j \mid y_i) \quad (6)$$

3. K-Nearest Neighbor

The K-Nearest Neighbor algorithm (Altman, 1992) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of

instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. 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 (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. 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 Yi, we sum the similarity of the neighbors of the same class. Then, the class Yi with the highest score is assigned to the data point by the K-Nearest Neighbors algorithm.

$$\hat{\Phi}(x') = argmax_{y_i \in c} \sum_{i=1}^{k} \delta(y_i, \Phi(x_i)) sim(x_i, x') \qquad (7)$$

4. Ensemble methods

Ensemble methods (Zhi-Hua, 2012) have rasied to be a major learning paradigm since the 1990s, with great efforts by two pieces of originating works. One is ensemble, it was found that predictions made by combination of sets of classifiers are more precise than predictions made by the best single classifier. A simplified illustration is depicted in Figure 1. The other is theoretical, in which it was proved that weak learners can become strong learners. Since strong learners are the goal, while weak learners are easy to obtain in real usage, this result opens a assuring direction of creating strong learners by ensemble methods.

Ensemble is composed of two steps. The first step is generating the base learners, the second step is combining those learners. It has commonly been assumed that the base learners should be as accurate as possible, and as diverse as possible.

This research proposed model is based on ensemble classification majority voting scheme over certain types of base classifiers which are of low computational complexity. It has used three base classifiers from different theoretical background to

avoid bias and redundancy. The three base classifiers are: 1) Naive Bayes, 2) K-Nearest Neighbor, and 3) decision tree.

Focus on Ensemble majority voting classification, each of the three base classifiers is an expert in a different region of the predictor space because they treat the attribute space under different theoretical basis (Alpaydin, 2010). The three classiiers could be combined in such a way to produce an ensemble majority voting classiier that is superior to any of the individual rules.Apopularway to combine these three base classiication rules is to let an ensemble classiier:

$$C(X) = mode\{h_1(X), h_2(X), h_3(X)\} \quad (8$$

to classify X to the class that receives the largest number of classiications (or votes). Mode is the value that appears most often in a set of data.

THAI SENTIMENT ANALYSIS FRAMEWORK

The objective of this research present sentiment analysis framework focused on ensemble classifiers methods to capture an emotional base on the opinion of the social media network dataset using natural language processing, specifically for only Thai language, which consists of four main steps:

1. The experiment using a collection of Thai Sentiment Text obtained from the Thai popular social media network webboard posts on internet including facebook twitter www.pantip.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. And then give to 3 expert readers are asked to label the best emotion that can be exposed from the text. There are six emotions category: anger, fear, joy, love, sadness, surprise. 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 SWATH (Smart Word Analysis for Thai) (Surapant, Paisarn & Boonserm, 1997) which is based on dictionary algorithm as a tokenizer in this Bag-Of-Words approach and identify Part-Of-Speech Tagging.

Thai sentiment dataset	CLASS
โง่เง่าเบาปัญญาสิ้นดี Too much Idiot.	ANGER
ขอบคุณที่เอามาแชร์ครับ น่ากลัวมาก ๆ จริง ๆ แท็กซี่สมัยนี้ Thanks for sharing, Now a day Taxis so scary.	FEAR
ดูคลิปนี้มาหลายปีละ กูก็ยังขำเหมือนเดิม 55 I have been watching this clip many years ago, It's always very funny LOL.	JOY
ชอบป๋ามาก ๆ แก่แล้วแต่ยังมีเสน่ห์เหลือล้น I love big dady so much. He's so attractive.	LOVE
บอกได้คำเดียว รับไม่ได้ ดูไม่ได้ น่าสงสารมาก ๆ Only one word unacceptable!! I can't watch. Too much poor.	SADNESS
โครตเจ๋งอ่ะ มีเพลงด้วยเหรอวะเดี๋ยวนี้ Cool!! Now a day, Have song like this	SURPRISE

2. 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. The output from this step will be used in the weighting scheme to assign the feature values as described in Section 2.

3. 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.

4. For classifying emotion step, I used weka (Hall et al., 2009). An open-source machine learning tool, to perform the experiments. I used the default settings for all algorithms. F For ensemble classifiers algorithm including Naive Bayes K-Nearest Neighbor, and Decision Tree classifier. The input comes from social media documents pre-classified into a set of emotion class. Figure 1 illustrates the thai sentiment analysis based on social media framework.

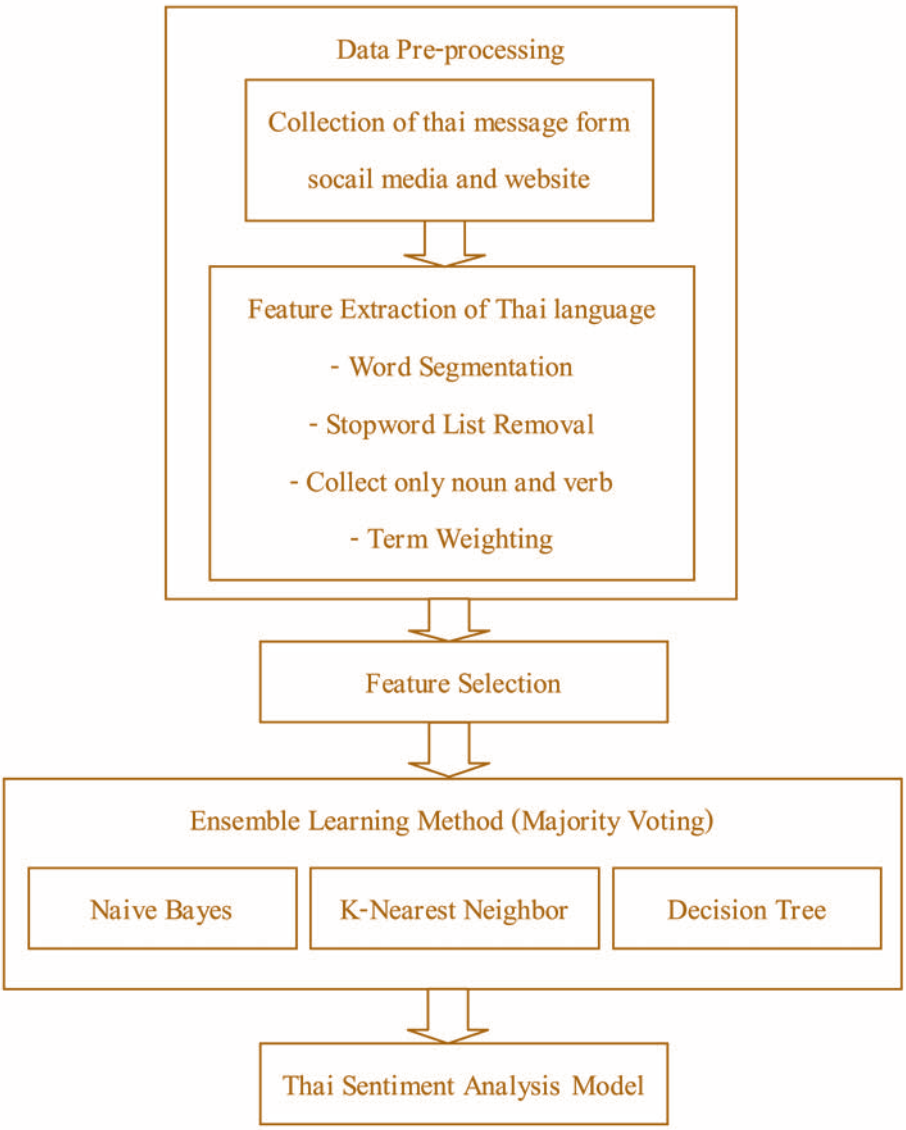

Figure 1. Thai sentiment analysis on social media using majority voting-based ensemble method framework

EXPERIMENT AND RESULTS

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix, Although a confusion matrix provides the information needed to determine how well a classification model performs,

summarizing this information with a single number would make it more convenient to compare the performance of different models. This can be done using a performance metric such as accuracy , which is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (9)$$

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 Sentiment Analysis based on social media dataset in combination with majority voting-based three learning algorithm are reported from Figure 2-5.

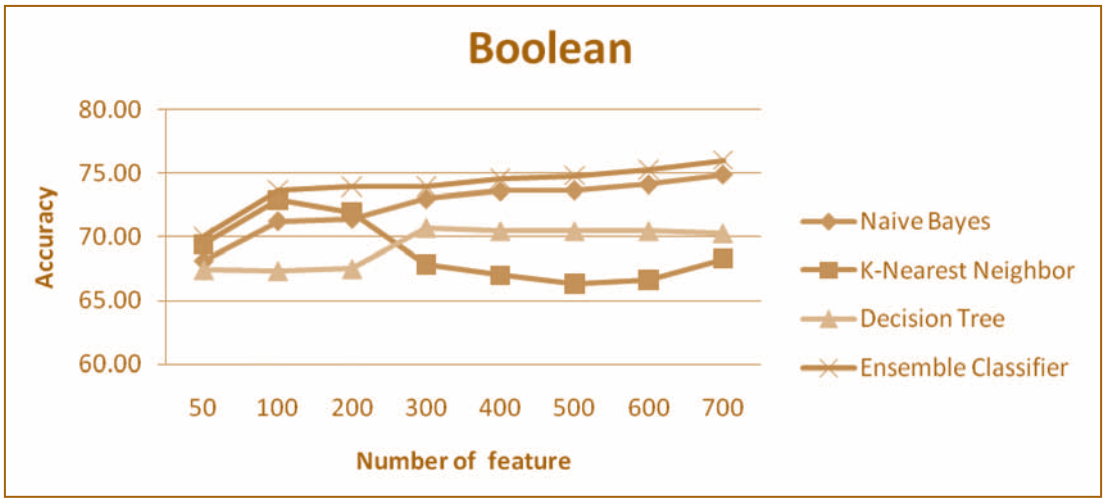

Figure 2. The experimental results from Boolean weighting

In figure 2 summarizes the results of classification with the Information gain feature selection method using four learning algorithms on Thai social media dataset after Boolean weighting. Five observations from Thai Sentiment Analysis based on Social Media Using Ensemble Classifiers Techniques were found. First, Ensemble Classifiers algorithm is the most accurate, followed by subordinate Naïve-Bayes, Decision Tree, and K-Nearest Neighbor algorithms respectively. Second, Ensemble Classifiers and Naïve-Bayes 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 Ensemble Classifiers algorithm at 76.04% with feature size of 700.

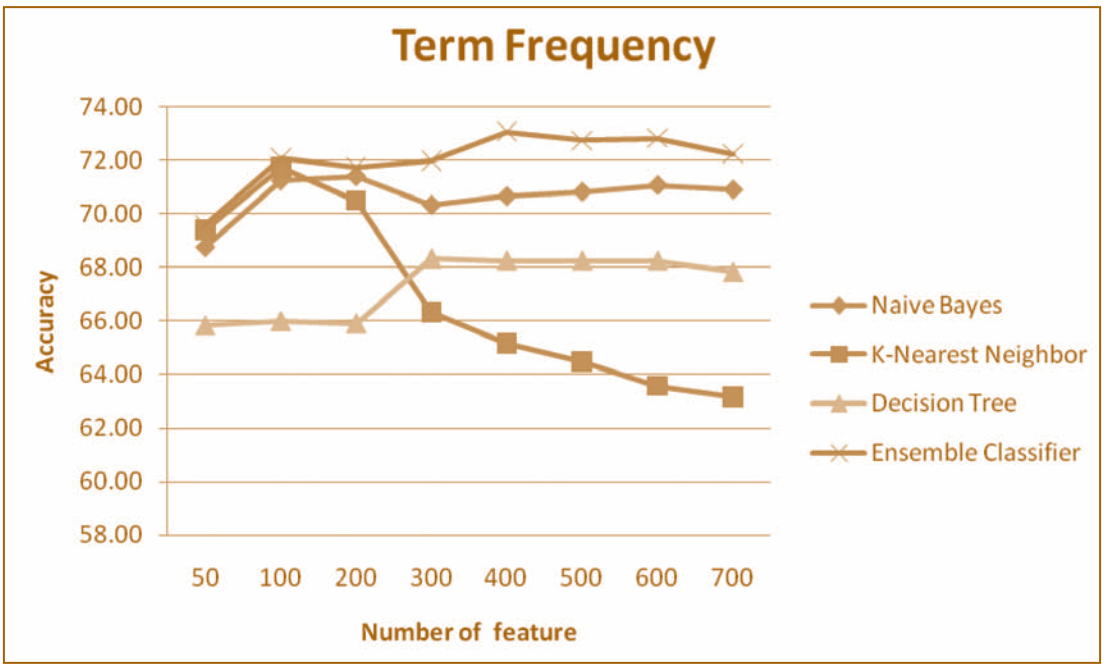

Figure 3. The experimental results from term frequency weighting

In figure 3 summarizes the results of classification with the Information gain using four learning algorithms on Thai social media dataset after Term Frequency weighting. Five observations from the Thai Sentiment Analysis were found. First, Ensemble Classifiers algorithm is the most accurate, followed by subordinate Naïve-Bayes, Decision Tree, and K-Nearest Neighbor algorithms respectively.

Second, Ensemble Classifiers and Naïve-Bayes have a similar trend of stable accuracy. It shows that the reduction of the feature does not affect the accuracy of Thai Sentiment Analysis. Third, Ensemble Classifiers and Decision Tree 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 Ensemble Classifiers algorithm at 73.08% with feature size of 400.

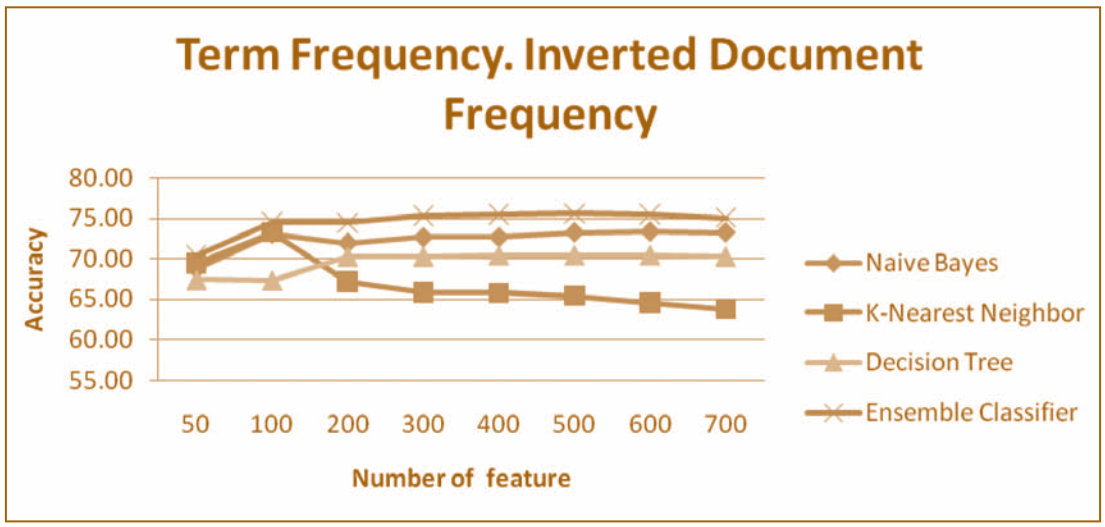

Figure 4. The experimental results from term frequency. Inverted document frequency weighting

In figure 4 summarizes the results of classification with the Information gain using four learning algorithms on Thai social media dataset after Term Frequency. Inverted Document Frequency weighting. Three observations from the Thai Sentiment Analysis were found. First, Ensemble Classifiers 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 Ensemble Classifiers algorithm at 75.60% with feature size of 500.

Table 1. The experimental results from comparison of weighting

Feature	TF Ensemble Classifier	TFIDF Ensemble Classifier	Boolean Ensemble Classifier
50	69.58	70.48	70.05
100	72.08	74.56	73.69
200	71.75	74.47	73.95
300	72.00	75.34	73.95
400	73.08	75.52	74.56
500	72.75	75.60	74.82
600	72.83	75.52	75.26
700	72.25	75.08	76.04

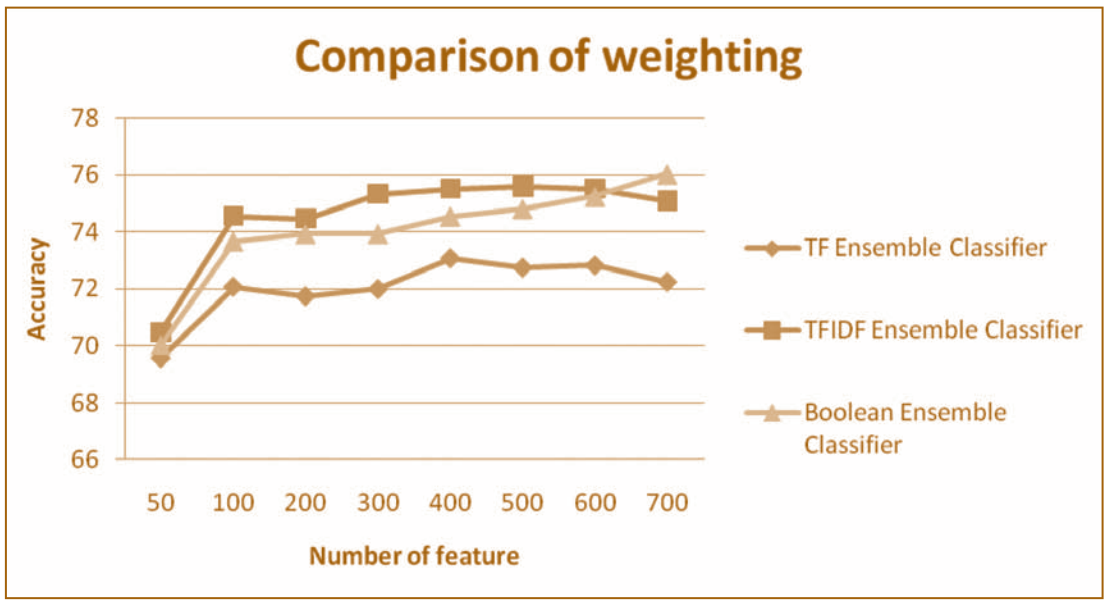

Figure 5. The experimental results from ensemble classifiers

In figure 5 summarizes the results of classification with the Information gain using Ensemble Classifiers algorithms on Thai social media dataset after Term weighting via Boolean, TF, TFIDF weighting, respectively. Four observations from the thai sentiment analysis using ensemble classifiers techniques were found. First, TFIDF weighting is more effective than another weighting with ensemble classifiers on thai sentiment analysis. Second, all term weighting schemes reached a maximum of accuracy point at the

full feature. Third, the best accuracy points base on Boolean weighting with ensemble classifiers were 76.04% at a feature size of 700. Finally, the TFIDF weighting is suitable for thai Sentiment Analysis on Social Media using ensemble classifiers techniques more than the other weighting.

CONCLUSIONS

In this research, I proposed Thai Sentiment Analysis on Social Media Using Majority Voting-Based Ensemble Method focusing on the comparison of various common term weighting schemes and multiple learning algorithms. I found TFIDF weighting with ensemble classifiers is most effective in our experiments. I also discovered that the TFIDF weighting is suitable for combination with the Information gain feature selection method. The TFIDF weighting with ensemble classifiers yielded the best performance with the accuracy over all algorithms. Based on our experiments, the ensemble classifiers algorithm with the Boolean weighting yielded the best performance with the accuracy of 76.04%. Our experimental results also reveal that feature weighting methods have a positive effect on the Thai Sentiment Analysis based on Social Media Framework.

REFERENCES

Alpaydin, E. (2010). ***Introduction to machine learning*** (2nd ed.). Cambridge, MA: MIT Press.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. ***The American Statistician. 46***(3), pp. 175-185.

Bartlett, P., & Shawe-Taylor, J. (1998). Generalization performance of support vector machines and other pattern classifiers. In Bernard Schölkopf, Christopher J. C. Burges & Alexander J. Smola (Eds.), ***Advances in Kernel methods - support vector learnin***g (pp. 43-54.). Cambridge, MA: MIT Press.

Bruzzone, L., Cossu, R., & Vernazza, G. (2004). Detection of land-cover transitions by combining multidate classifiers. ***Pattern Recognition Letters, 25***(13), pp. 1491-1500.

Buhlmann, P., & Yu, B. (2003). Boosting with L2 loss: Regression and classification. ***Journal of the American Statistical Association, 98,*** pp. 324-338.

Dietterich, T. G. (2000). an experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. ***Machine Learning, 40,*** pp. 139-157.

_____. (2001). Ensemble methods in machine learning. In Josef Kittler & Fabio Roli (Eds.), ***Multiple classifier systems. Proceeding of First International Workshop, MCS 2000 Lecture notes in computer science, vol. 1857*** (pp. 1-15). Cagliari, Italy: Springer.

Dimitriadou, E., Weingessel, A., & Hornik, K. (2003). A cluster ensembles framework. In Ajith Abraham, Mario Köppen & Katrin Franke (Eds.), ***Design and application of hybrid intelligent systems*** (pp. 528-534). Amsterdam, Netherlands: IOS Press.

Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian ensemble learning. ***Decision Support Systems, 68,*** pp. 26-38.

Gaikwad, D. P., & Thool, R. C. (2015). Intrusion detection system using bagging ensemble method of machine learning. In Juan E. Guerrero (Ed.), ***Proceedings of First International Conference on Computing Communication Control and Automation ICCUBEA 2015*** (pp. 291-295). Pimpri-Chinchwad, India: College of Engineering.

Geman, S., Bienenstock, E., & Doursat, R. (1995). Neural networks and the bias variance dilemma. ***Neural Computation, 4,*** pp. 1-58.

Hall, et al. (2009). The WEKA data mining software: An update. ***SIGKDD Explorations, 11***(1), pp. 10-18.

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural networks, and genetic algorithm: A case study in romantic decision support. ***Decision Support Systems, 32***(4), pp. 361-377.

Mangiameli, P., West, D., & Rampal, R. (2004). Model selection for medical diagnosis decision support systems. ***Decision Support Systems, 3***6(3), pp. 247-259.

Nivet, Chirawichitchai. (2015). Developing term weighting scheme based on term occurrence ratio for sentiment analysis. In Kuinam J. Kim (Ed.), ***Lecture notes in electrical engineering volume 339*** (pp. 737-744). Berlin, Germany: Springer-Verlag Berlin Heidelberg.

Polettini, N. (2004). ***The vector space model in information retrieval - term weighting problem.*** Trento, Italy: Department of Information and Communication Technology, University of Trento.

Quinlan, J. R. (1987). Simplifying decision trees. ***International Journal of Man-Machine Studies, 27***(3), pp. 221-234.

Rish, I. (2001). An empirical study of the Naive Bayes classifier. In ***Proceedings of IJCAI-2001 workshop on Empirical Methods in AI (also, IBM Technical Report RC22230)*** (pp. 41-46). New York, NY: IBM.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. ***LMagazine Communications of the ACM, 18***(11), pp. 613-620.

Surapant, Meknavin, Paisarn, Charoenpornsawat, & Boonserm, Kijsirikul. (1997). Feature-based Thai word segmentation. In ***Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97)*** (pp. 35-46). Bangkok, Thailand: National Electronics and Computer Technology Center

Tan, A. C., Gilbert, D., & Deville, Y. (2003). Multi-class protein fold classification using a new ensemblemachine learning approach. ***Genome Informatics, 14,*** pp. 206-217.

Tumer, K., & Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. ***Connection Science, 8***(3-4), pp. 385-404.

Yang, Y., & Pedersen, J. P. (1997). A comparative study on feature selection in text categorization. In Douglas H. Fisher (Ed.), ***Proceedings of the Fourteenth International Conference on Machine Learning ICML '97*** (pp. 412-420). San Francisco, CA: Morgan Kaufmann Publishers.

Zhi-Hua, Zhou. (2012). ***Ensemble methods: Foundations and algorithms.*** Boca Raton, FL: Chapman and Hall/CRC.