ICIC Express Letters
Volume **13**, Number **3**, March **2019**

ICIC International ©2019 ISSN 1881-803X
pp. **217–224**

SENTIMENT ANALYSIS OF THE BURMESE LANGUAGE USING n-GRAM-BASED WORDS

Myat Lay Phyu[1] and Kiyota Hashimoto[2]

[1]College of Computing
[2]Interdisciplinary Graduate School of Earth System Science
and Andaman Natural Disaster Management (ESSAND)
Prince of Songkla University
80 Moo 1, Vichitsongkram Road, Kathu, Phuket 83120, Thailand
{ myatlayphyu096; kiyota.hashimoto }@gmail.com

Received August 2018; accepted November 2018

Abstract. *Text analysis has been an important research area. Among text processing, opinion mining which is deciding the opinions and giving feelings of others is popular. The researchers have proposed different ways to give the opinions of people automatically. In this case, low resource languages are still difficult to treat due to the unavailability of annotated big corpora and basic natural language processing tools. This research proposes a new method to use a character-based variable-length n-gram word model, which makes n vary within the pre-set thresholds, and select frequent strings of each n as n-length words. We employed this method for words segmentation and the sentiment values are calculated based on variable-length of n-gram-based words. Finally, sentiment analysis of Burmese news articles is processed whether the news is positive or negative, and achieved a similar result with Conditional Random Field (CRF) based ordinary word segmentation with a small size of supervised data. This enables to treat low resource languages without focusing on language specific characteristics.*

Keywords: Sentiment analysis, The Burmese language, Variable-length n-gram word, CRF

1. **Introduction.** Sentiment analysis has been extensively explored to estimate the sentiment value, positive or negative, of opinions, reviews, and many other types of texts [1,2]. Typically its target is text, and thus sentiment analysis is usually based on various pre-processing including sentence segmentation, word segmentation, part-of-speech tagging, word sense disambiguation, etc., as well as appropriate sizes of class-labeled datasets with which supervised machine learning training is conducted. In the case of languages such as English, Spanish, and Japanese, such tools and data resources are readily available. In the case of languages such as Burmese, Khmer, and Lao, to name just a few of so-called low resource languages, they are often readily unavailable. This is all the more critical when the target language is written without explicit word delimiters; in other words, when words are written concatenated without breaks with spaces or other punctuations. The Burmese language is exactly of this type. Currently a word segmentation tool of a sufficient quality and a stardard corpus is unavailable, and the Burmese writing system concatenates words, though spaces are often used to separate phrases. There are roughly two directions to cope with these difficulties. One is, naturally expected, to prepare necessary preprocessing tools and datasets for the target language, which requires time, cost, and efforts, which are also not readily available in most cases. The other is to develop an alternative way to do without those. There are some studies trying some language tasks without assuming word segmentation [3], but there is no such study for

DOI: 10.24507/icicel.13.03.217

the Burmese language. As Myanmar began developing fast, practical methods that do not assume costly preparations are indeed desirable.

In this paper, we propose a new method to conduct sentiment analysis with variable-length n-gram words for the Burmese language. The overview of our research procedure is shown in Figure 1. n-gram word is defined as a string of characters of n. Thus, n-gram words will not be a good substitute for ordinary words when n is fixed. Our proposal is to make n vary within the pre-set thresholds. This method begins with the maximum $n_{\max}$ to pick up frequent strings of n-length, and then proceeds pick-ups from the rest of the text iteratively to the minimum $n_{\min}$. We evaluate this proposal with the Burmese language, one of the low resource languages, in comparison with an ordinary word segmentation model that were previously proposed by us with a tiny set of data [7]. With these variable-length n-gram words, sentiment classification is conducted with two feature sets: *tf-idf* of each word and word sentiment values that are calculated with SO-LSA (Semantic Orientation-Latent Semantic Analysis) [4] method. The sentiment classification is done with Support Vector Machines (SVM) [5]. The experimental results show that our variable-length n-gram word-based sentiment analysis achieved better results compared to ordinary word-based sentiment analysis. Importantly, our method does not refer to linguistic characteristics specific to the target language, and thus this method can be used for other languages without much time, cost, and effort to prepare pre-processing tools.

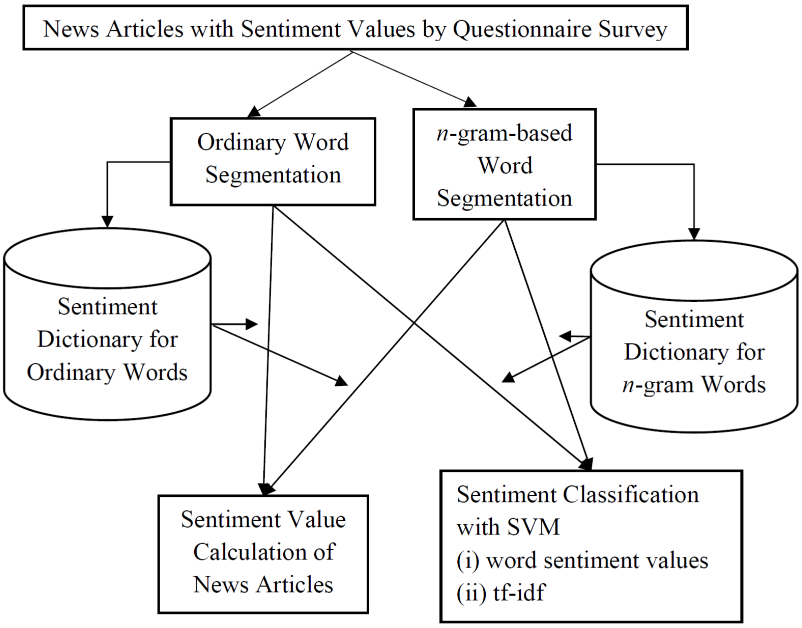
News Articles with Sentiment Values by Questionnaire Survey
Ordinary Word Segmentation
n-gram-based Word Segmentation
Sentiment Dictionary for Ordinary Words
Sentiment Dictionary for n-gram Words
Sentiment Value Calculation of News Articles
Sentiment Classification with SVM
(i) word sentiment values
(ii) tf-idf

Figure 1. Overview of research procedure

The organization of this paper consists of 7 sections: the Burmese language, related works, word segmentation with Conditional Random Fields (CRF) – a baseline model, variable-length n-gram words, sentiment analysis of Burmese newspaper articles, and conclusion.

2. **The Burmese Language.** The Burmese language is the official language of the Republic of the Union of Myanmar. The Burmese writing system consists of 33 main alphabets or consonants (some people say 34 because one consonant has two types), characters such as 7 independent vowels, 7 dependent vowels, 4 medials, 2 final symbols, 2 tone marks, 4 abbreviations, 2 types of punctuation and numerals. The Burmese text is written from left to right. It requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability. However, there are no specific rules for space adding. Let us illustrate Burmese writing in Figure 2.

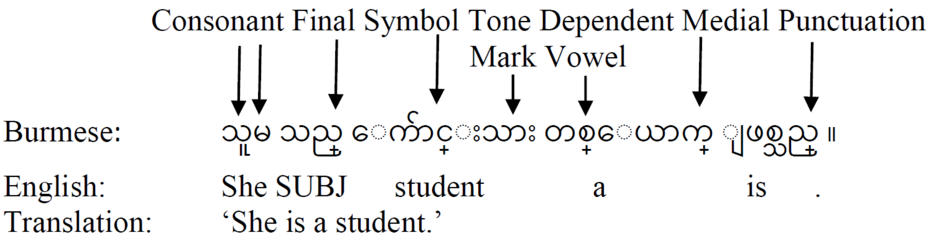

Figure 2. An example of Burmese sentence with English translation

3. **Related Works.** Turney and Littman introduced the calculation of Semantic Orientation (SO) of a word from its associated statistical information with a set of positive and negative seed words [4]. If the value of semantic orientation of a word is positive, the word is positive; otherwise, it is negative. This approach is evaluated in two ways based on two different statistical measures of word association called Pointwise Mutual Information (PMI), which measures the association degree between two words by the frequency of their co-occurrence, and Latent Semantic Analysis (LSA), which measures the association degree of two words by comparing the contexts in which they co-occur. First to calculate the semantic orientation of a word, Latent Semantic Analysis (LSA) is calculated. Calculation of LSA includes two main parts. First *tf-idf* (term frequency – inverse document frequency) matrix (the original matrix X) is constructed. In the matrix, the row vectors represent the words and the column vectors represent the documents (news articles). Each cell represents the weight, *tf-idf* score, of the corresponding word in the corresponding document. The next step is to apply Singular Value Decomposition (SVD) to compressing the original matrix, X. SVD decomposes X into a product of three matrices, $X = UDV^{\mathrm{T}}$, where U and V are orthogonal matrices and D is a diagonal matrix of singular values. Let D_k, the diagonal matrix with the top k singular values, U_k and V_k be formed by selecting the corresponding columns from U and V. The matrix $U_k D_k V_k^{\mathrm{T}}$ can be considered as a compressed version of the original matrix.

The similarity of two words, LSA (word1, word2), is measured by calculating the cosine similarity of two corresponding vectors in the compressed matrix. Semantic Orientation, SO-LSA of a word is calculated based on the following equation:

$$\begin{aligned}\text{SO-LSA (word)} = {} & \textstyle\sum \text{pword} \in \text{Pwords LSA (word, pword)} \\ & - \textstyle\sum \text{nword} \in \text{Nwords LSA (word, nword)}\end{aligned} \tag{1}$$

Here, Pwords and Nwords mean the set of positive and negative seed words. If SO-LSA (word) value is positive, the word is classified as positive word and it is negative word when SO-LSA (word) value is negative. In this way, a sentiment dictionary, consisting of entries and their sentiment values, is created according to the target data.

Their experiments suggested that SO-LSA method used the data more efficiently than SO-PMI method and SO-LSA might provide better accuracy than SO-PMI for a corpus of comparable size. By obtaining the sentiment value of each word, a sentiment dictionary can be created according to the target data. In this study, we followed their method, SO-LSA, to calculate the sentiment values of the word.

When we have sentiment values of words, the sentiment value of a sentence or a text can be estimated as the summation value of all words in it [1]. If the resulted value is positive, the article is classified as 'positive'. Otherwise, it is classified as 'negative'. The formula is as follows:

$$\sum_{i=1}^{i=n} V_{w_i} \tag{2}$$

where w_i is the i-th word in a news article and V_{w_i} is the sentiment value of w_i. This method is employed particularly when the ground truth data is not available.

Another way to estimate the sentiment value of a sentence or a text is to use a machine learning classification approach. In this approach, we need the ground truth class labels, and word sentiment values, or some other values including *tf-idf* values, can be used as a set of features. Many studies have been conducted with many different machine learning techniques, and one of the best methods is Support Vector Machines (SVM). SVM is one of the most popular and efficient machine learning algorithms [5]. In this paper, we employ SVM with a linear kernel using scikit-learn with default parameter setting except for C value which is 0.1 or 1.0. A linear kernel is selected because it is considered to be most suitable kernel for text analysis [9].

4. Word Segmentation with Conditional Random Fields (CRF) – A Baseline Model.

In our previous study [7], a way to process the word segmentation for the Burmese language with CRF framework was proposed, which is the current state-of-the-art model and is employed here as a baseline model to evaluate our proposal in this paper.

A machine learning framework, CRF, for building probabilistic undirected graphical models to segment and label the structured data was proposed and [6] reported the performance of CRF outperforms the previous models on natural language data. The largest benefit of CRF is that it can handle the various types of non-independent features of input. The sequence of input data to the CRF is structured. CRF learns the structured data and constructs a conditional model. The model predicts the probabilities of possible label sequence of input data based on the previously observed data.

In [7], characters clustering is employed in which character clusters mean groups of some inseparable characters due to language characteristics. For characters clustering, a set of 29 types of Burmese Character Clusters (BCCs), in which 22 of them corresponds to equivalents in Thai Character Cluster (TCC) from [8] and seven of them are newly proposed as rules. CRF is applied as a sequential labeling machine learning method. The sequence of input data is labeled as ‘B’ for the beginning character cluster of a word, ‘I’ for the inner or inside character cluster of a word, ‘E’ for the end character cluster of a word and ‘S’ is the single character cluster word. The experiment was tested on a small set of data. The data set contains only 5 news articles which contain approximately 8,000 single words in a total of 191 sentences. Some single words are grouped to form compound words and then the data remains nearly 3,500 words. The experiment achieved the accuracy of 98.8%.

Importantly, as CRF is a supervised machine learning method, its training needs labeled data, which are not easily available in many low resource languages. In [7], a tiny set of labeled data was manually prepared. Such preparation requires tremendous time and efforts, also considering another fact that linguistic agreements on word and grammar are also not fixed in many low resource languages. Thus, it is often realistically difficult to prepare labeled datasets of similar quantity and quality to some major languages such as English, which is why variable-length n-gram-based word segmentation is proposed in this paper.

5. Variable-Length n-gram-based Word Segmentation.

In the previous section, we introduced the ordinary word segmentation. As noted in the last paragraph of the previous section, the effect of the ordinary word segmentation has a severe limitation in most low resource languages because of the lack of a large size of labeled corpora, or datasets.

Instead of ordinary word segmentation, an n-gram word model has often been employed. An n-gram word model defines a word as a string of characters of n-length, and thus all the n-gram words in a dataset are error-free by definition. However, when n is fixed, those n-gram words are far from ordinary words, and the number of n-gram words is usually much larger than the number of the ordinary words in the same dataset. In order to reduce these issues, we propose a variable-length n-gram-based word segmentation. A variable-length n-gram-based word segmentation varies n. According to manual observations of the target language, the maximum $n_{\max}$ and the minimum $n_{\min}$ is set, which is the only part to consult the language characteristics of the target language. Then, n-gram word pick-up is conducted first from $n_{\max}$ by choosing sufficiently frequent $n_{\max}$-gram words. Every selected $n_{\max}$-gram word is rewritten as $X_{\max}$ in the dataset. The next step chooses $(n_{\max}-1)$-gram words of sufficient frequency, and they are rewritten as $X_{\max-1}$. The same procedure is repeated until $n_{\min}$. With this iterative procedure, selected n-gram words are more similar to ordinary words and the total number of selected n-gram words is much smaller than the number of n-gram words of a fixed length. In this paper, according to the characteristics of the Burmese writing system as well as Burmese words, $n_{\max}$ is set to be 25 and $n_{\min}$ to be 3. Let us illustrate this procedure in Figure 3. For a notational convenience, $X_{\max}$ is P, $X_{\max-1}$ is O, ... and $X_{\min}$ is 3.

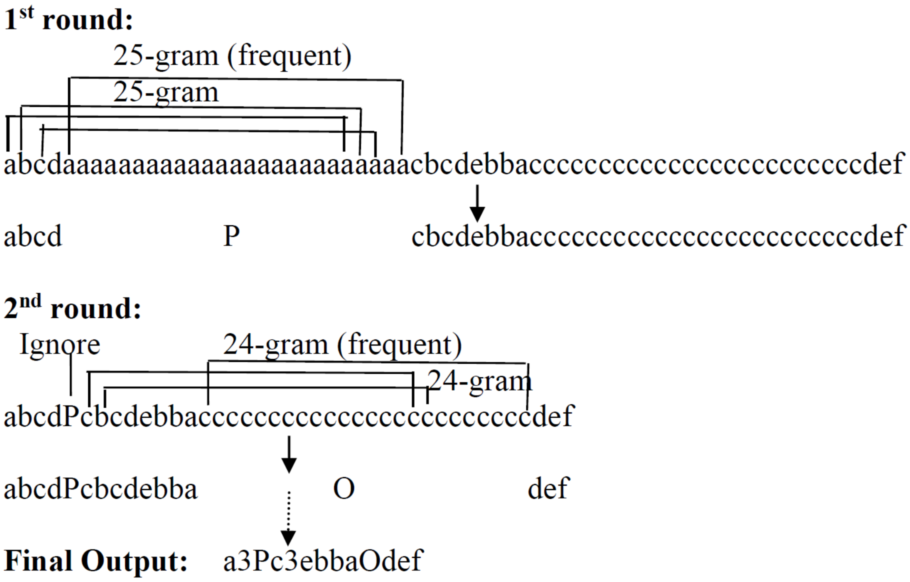
1st round:
25-gram (frequent)
25-gram
abcdaaaaaaaaaaaaaaaaaaaaaaaaacbcdebbaccccccccccccccccccccccccccdef
abcd
P
cbcdebbaccccccccccccccccccccccccccdef
2nd round:
Ignore
24-gram (frequent)
24-gram
abcdPcbcdebbacccccccccccccccccccccccccdef
abcdPcbcdebba
O
def
Final Output:
a3Pc3ebbaOdef

Figure 3. Example of variable-length n-gram word string conversion

6. Sentiment Analysis of Burmese Newspaper Articles.

6.1. **Data description.** Newspaper articles of 7Day Daily, one of the most popular private newspapers in Myanmar are used. The daily news of newspaper can be obtained from its official website [7]. Among several topical sections, we collected 1,280 news only from the Opinion group. This whole dataset is used for data processing steps. For classification, we made surveys for 500 news articles to get the opinions from the Burmese people whether the news is positive or negative, to prepare class labels. Among 500 news, the number of positive news is 226, negative news is 211 and neutral news is 63, and 437 news (positive and negative news among 500 news) is used for classification.

6.2. **Experiment setting.** We compare two ways to perform sentiment analysis of the Burmese news articles. The general setting for both approaches is the same, summarized in Table 1.

In Experiment I, the dataset contains 1,280 news articles which contain 1,008,274 ordinary words, which is segmented into words with the CRF-based method. After removing stop words, which is a manually prepared set of 278 words, only the words of more than five occurrences are used. The extracted data from 1,280 news contains 642,887 words.

Table 1. Preprocessing part of experiments

Experiment I	Experiment II
1. Use all news articles (1,280 news) 2. Replace English punctuations !"#$%&'()*+,-./:;<=>?@[¥]^_`{\|} ~ with a 'space' 3. Usual Word Segmentation with CRF (Section 4) 4. Remove English words and numbers (both Burmese and English numbers) 5. Replace Burmese punctuation: a comma with a 'space' 6. Replace Burmese punctuation: a full stop with a '$' 7. Remove stop words	1. Concatenate all news articles (1,280 news) and insert 'Ⓡ' at the end of each news 2. Replace English punctuations !"#$%&'()*+,-./:;<=>?@[¥]^_`{\|}~ with 'U' 3. Replace English words with 'W' 4. Replace numbers (both Burmese and English numbers) with 'X' 5. Replace Burmese punctuations: comma and full stop with '$' 6. *n*-gram-based word segmentation (Section 5) 7. Remove *n*-gram words which have very high frequency

Experiment II uses the same dataset as the Experiment I. During variable-length *n*-gram-based word segmentation, infrequent *n*-gram words are removed. After the segmentation process, very high frequent *n*-gram words are also removed as potential stop words. After that, 1,280 news data contain 894,502 *n*-gram words.

The same seed word list is applied in both cases. The chosen seed words are as follows:

Positive: ['ကောင်းသော', 'ဖွံ့ဖြိုး', 'တိုးတက်ရေး', 'ပျော်ရွှင်သော', 'အောင်မြင်', 'အဆင်ပြေ', 'မှန်ကန်သော', 'ငြိမ်းချမ်းရေး']

Negative: ['ကျဆင်း', 'အဆင်မပြေ', 'ပြဿနာ', 'ဆိုးသော', 'အခက်အခဲ', 'အန္တရာယ်', 'ဝေဖန်မှု', 'ဒုက္ခ']

The translation of these seed words is:

Positive: ['good', 'develop', 'improvement', 'happy', 'success', 'convenient', 'correct', 'peace']

Negative: ['decline', 'inconvenient', 'problem', 'bad', 'difficulty', 'danger', 'criticism', 'trouble']

Note that in Experiment II, an *n*-gram word containing a seed word but not containing a negative word is also regarded as the same seed word. The above seed words lists are used for the next step, calculating the sentiment values and constructing sentiment dictionary. For this step, LSA and SO-LSA methods are performed to calculate the sentiment values and construct sentiment dictionary. The classification for each news is conducted in two ways: Summation of Sentiment Values and Support Vector Machine (SVM), which is explained in Section 3. For SVM, we compared two features: *tf-idf* value and sentiment value.

Experiment II is conducted with 10-fold cross-validation.

6.3. **Result and discussion.** Table 2 shows the original data set and extracted data set in each step. For the classification, 500 articles, among 1,280 articles, are used. These 500 articles are manually prepared for class labels, positive or negative, based on the questionnaire survey of 100 Burmese people.

Tables 3 and 4 show the performance of the experiment in two different ways of classification. The results of sentiment value summation, the result using ordinary words is better than with *n*-gram words in all four metrics, as shown in Table 3. This is easily expected because the simple sentiment value summation of all sentiment words in each article can use a more correct set of sentiment words in the case of ordinary words while the case with *n*-gram words uses them in a more indirect way. On the other hand, when

Table 2. Data description

Number of extracted words in 500 articles	Usual words with CRF (Experiment I)	n-gram words (Experiment II)
Total words	115,144	223,437
Unique words	8,980	78,648

Table 3. Classification with sentiment values summation

	Accuracy	Precision	Recall	F-measure
Experiment I (Ordinary words)	**73.23%**	**72.80%**	**76.99%**	**74.84%**
Experiment II (n-gram words)	72.08%	76.00%	67.26%	71.36%

Table 4. Classification with SVM

	SVM	Accuracy	Precision	Recall	F-measure
Experiment I (Ordinary words)	*tf-idf* values	74.59%	**76.11%**	74.44%	75.27%
	Sentiment Values	83.04%	84.35%	**83.72%**	84.04%
Experiment II (n-gram words)	*tf-idf* values	**75.72%**	74.57%	**82.16%**	**78.18%**
	Sentiment Values	**84.22%**	**87.50%**	81.76%	**84.53%**

SVM is applied for classification, the result tells that the n-gram word segmentation model achieved a better performance than the ordinary word segmentation model with accuracy, precision, and F-measure, as in Table 4. The result also shows that the use of sentiment value of n-gram words as features achieved better results than *tf-idf* values. One of the reasons of this result is that our ordinary word segmentation model is based on our tiny dataset to train a CRF-based word segmentation classifier, and it is likely to have more errors for the 1,280 articles dataset. Note that if one tries to improve this situation, one has to take much larger time, cost, and efforts to prepare a larger dataset to train a CRF-based word segmentation classifier. On the other hand, our variable-length n-gram-based word segmentation seems to select many n-gram words that are quite similar to ordinary words. Thus, we conclude that our proposed method is a better option for low resource languages both in classification quality and in efficiency.

7. **Conclusion.** In this paper, we proposed a method to obtain variable-length n-gram words that can be used as similar to ordinary words and evaluated this method for sentiment analysis. Our results show that our method achieved better results compared to the ordinary word segmentation that is constructed with a tiny set of manually tagged dataset, which indicates that variable-length n-gram word model is a promising method for low resource languages because our method does not need a large annotated data or elaborately developed word segmentation tools, and it can be applied to other languages fairly easily.

However, the number of n-gram words that are obtained is large, which means that many n-words that are far from ordinary words seem to be major obstacles. Therefore, one of the future works is how to reduce the number of n-gram words and to improve the quality of the obtained n-gram word set.

Acknowledgment. This work was supported by the Higher Education Research Promotion and the Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission.

REFERENCES

[1] P. D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.417-424, 2002.

[2] C. Quan and F. Ren, Target based review classification for fine-grained sentiment analysis, *International Journal of Innovative Computing, Information and Control*, vol.10, no.1, pp.257-268, 2014.

[3] B. P. King, *Practical Natural Language Processing for Low-Resource Language*, Ph.D. Thesis, University of Michigan, 2015.

[4] P. D. Turney and M. L. Littman, Measuring praise and criticism: Inference of semantic orientation from association, *ACM Trans. Information Systems*, vol.21, no.4, pp.315-346, 2003.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1998.

[6] J. Lafferty, A. Mccallum and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. of the 18th ICML*, pp.282-289, 2001.

[7] M. L. Phyu and K. Hashimoto, Burmese word segmentation with character clustering and CRFs, *Proc. of the 14th JCSSE*, pp.1-6, 2017.

[8] T. Thanaruk, V. Sornlertlamvanich, T. Tanhermhong and W. Chinnan, Character cluster based Thai information retrieval, *Proc. of IRAL'00*, pp.75-80, 2000.

[9] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, *Machine Learning: ECML-98*, pp.137-142, 1998.

[10] *7Day Daily*, http://www.7daydaily.com/.