Automatic Myanmar News Classification

Khin Thandar Nwet, Aye Hninn Khine, Khin Mar Soe

University of Computer Studies, Yangon

{ khinthandarnwet, ayehninnkhine, khinmarsoe}@ucsy.edu.mm

Abstract

Text classification is one of the major tasks of natural language processing and included in the interesting research areas of text data mining, which is about looking for patterns in natural language text. This paper applies two well-known classification algorithms. Algorithms applied are Naïve Bayes and k-Nearest Neighbors (KNN). These well-known algorithms are applied on collected Myanmar News dataset. Dataset used consists from 1200 documents belongs to 4 categories. The goal of text classification is to classify documents into a certain number of pre-defined categories. News corpus is used for training and testing purpose of the classifier. Feature selection algorithm is used in the proposed system to select the most relevant features from training data. Results show that precision and recall values using k-NN is better than Naïve Bayes. This research makes a comparative study between mentioned algorithms.

Keywords: Text Classification, Natural Language Processing, Naïve Bayes, k-Nearest Neighbors classifier

1. Introduction

With the rapid growth of the internet, the availability of on-line text information has been considerably increased. As a result, text mining has become the key technique for handling and organizing text data and extracting relevant information from massive amount of text data. Text mining may be defined as the process of analyzing text to extract information from it for particular purposes, for example, information extraction and information retrieval. Typical text mining tasks include text classification, text clustering, entity extraction, and sentiment analysis and text summarization.

Text classification is involved in many applications like text filtering, document organization, classification of news stories, searching for interesting information on the web, spam e-mail filtering etc. These are language specific systems mostly designed for English, European and other Asian languages but very less work has been done for Myanmar language. Therefore, developing classification systems for Myanmar documents is a challenging task due to morphological richness and scarcity of resources of the language like automatic tools for tokenization, feature selection and stemming etc.

In this paper, an automatic text classification system for Myanmar news is proposed. The proposed system is implemented by using supervised learning approach which defines as assigning the pre-defined category labels to the text documents based on the likelihood suggested by the training set of labeled documents. For the training data set, news from Myanmar media websites are manually collected, labeled and stored in the training data set. Chi Square

function is used as a feature selection method and Naïve Bayes theory is applied in implementing the text classifier. In this Naïve Byes text classifier, word frequencies are used as features.

This paper used k-Nearest Neighbors and Naïve Bayes Classifier in Myanmar text classification. KNN and Naïve Bayes are two classifiers frequently used for Text Classification, and they are similarity based and probability based. This paper aims at making a comparative study between mentioned algorithms on a Myanmar data set. Actually and up to this paper date authors did not find any research that make a comparative study between mentioned algorithms on a Myanmar Language.

The remaining parts of this paper are organized as follows. In section 2, the nature of Myanmar language is discussed. The related works are explained in section 3. Then, the overview of the proposed system is shown in section 4 and the step by step Naïve Bayes algorithm for text classification is illustrated in section 5. In section 6, K Nearest Neighbors is discussed. In the later sections, section 7 and section 8, the experimental work is described and the paper is concluded specifically.

2. Myanmar Language

Myanmar language is an official language of the Republic of the Union of Myanmar. It is written from left to right and no spaces between words, although informal writing often contains spaces after each clause. It is syllabic alphabet and written in circular shape. It has sentence boundary marker. It is a free-word-order and verb final language, which usually follows the subject-object-verb (SOV) order. In particular, preposition adjunctions can appear in several different places of the sentence. Myanmar language users normally use space as they see fit, some write with no space at all. There is no fixed rule for word segmentation. Word segmentation is an essential task in preprocessing stage for text mining processes. Many researchers have been carried out Myanmar word segmentation in many ways including both supervised and unsupervised learning. We use Myanmar Word Segmentation with Syllable level Longest Matching approach in the proposed system.

3. Related Work

Many algorithms have been applied for Automatic Text Categorization. Most studies have been devoted to English and other Latin languages. However, very few researches have been carried out on Myanmar text.

In the paper written by Young joong Ko and Jungyun Seo, unsupervised learning method was used for Korean Language text classification. The training documents were created automatically using similarity measurement and Naïve Bayes algorithm was implemented as the text classifier [1].

S Kamruzzaman and Chowdhury Mofizur Rahman from United International University (Bangladesh) has been developed text classification tool using supervised approach. The Apriori algorithm was applied to derive feature set from pre-classified training documents and Naïve Bayes classifier was then used on derived features

for final categorization [2]. And, CheeHong Chan, Aixin Sun and Ee-Peng Lim presented classifier to classify news articles from the Channel News Asia [4]. The unique feature of this classifier was that it allowed users to create and maintain their personalized categories. Users can create their personalized news category by specifying a few keywords associated with it. These keywords were known as the category profile for the newly created category. To extract feature words from the training dataset, TfIDf (term frequency–inverse document frequency) algorithm is used in this classifier. Support Vector Machine is applied in the classification stage of this classifier.

Riyad al-Shalabi implemented KNN on Arabic data set. He has reached 0.95 micro-average precision and recall stores. Also he uses 621 Arabic text documents belong to six different categories. He has used a feature set consist of 305 keywords and another one of 202 keywords. Selection of keywords based on Document Frequency threshold (DF) method [13].

There are many supervised learning algorithms that have been applied to the area of text classification, using pre-classified training document sets. Those algorithms, that used classification, include K-Nearest Neighbors (K-NN) classifier, Naïve Bayes (NB), decision trees, rocchio's algorithm, Support Vector Machines (SVM) and Neural Networks [13, 14].

4. Overview of the Proposed System

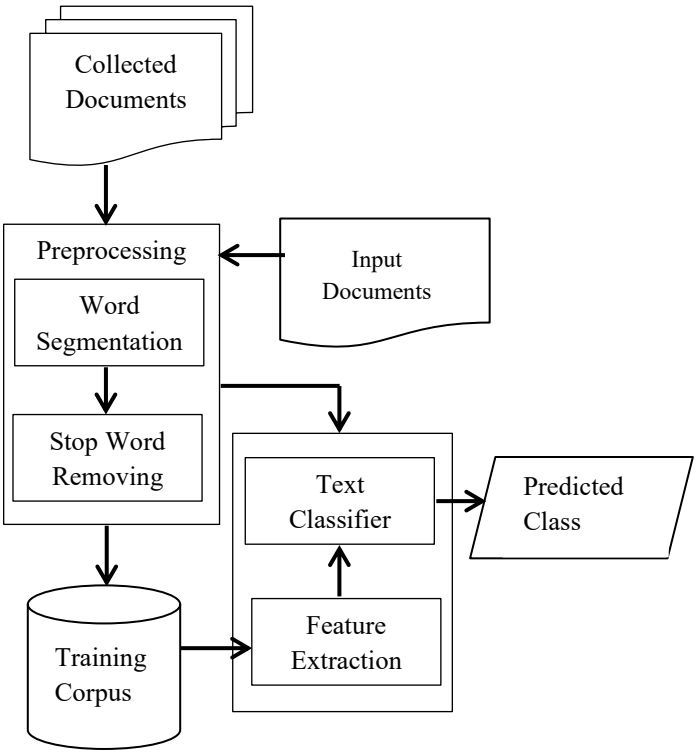

Figure 1 Proposed System Design

As shown in figure 1, there are two phases in the proposed system, namely, training and testing phases. In the training phase, preprocessing of collected raw data and feature selection are carried out. In the testing phase, feature words from training corpus are extracted and later the classification process is done. In the proposed system, four categories such as politic, business, sport and entertainment are defined.

Since the proposed system uses supervised learning approach, it needs to collect raw data to create training corpus. Therefore, news from Myanmar media websites such as news-eleven.com,

7daydaily.com and popularmyanmar.com are manually collected for each category.

4.1 Preprocessing

Preprocessing plays an important role in text mining techniques and applications. It is the first step in the text mining process. Preprocessing step is crucial in determining the quality of the classification stage. It is important to select the significant keywords that carry the meaning and discard the words that do not contribute to distinguishing between the documents. The first part of the text classification is syllable and word segmentation. Space delimited languages are easy to tokenize or segment words in the sentence. However, some language such Japanese, Chinese, Thai, India and Myanmar are not clearly delimited by a space. Thus, syllable segmentation is essentially needed as text preprocessing step. First, sentences in collected documents are segmented into words using longest matching approach. And then, stop words, punctuations and special characters are removed from the collected documents. Myanmar stop words are တွင်၊ မယ်၊ များ၊ သည်၊ သည့်၊ များ etc. These words have very low discrimination value, since they occur in every Myanmar document. Hence they do not help in distinguishing between documents with contents that are about different topics.

4.2 Feature Selection

Feature selection is the process of selecting a subset of the words occurring in the training set and using only this subset as features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. This is of particular importance for classifiers. Second, feature selection often increases classification accuracy by eliminating noise features. A noise feature increases the classification error on new data. The size of the vocabulary used in the experiment is selected by choosing words according to their Chi Square ($\boldsymbol{\chi^2}$) statistic with respect to the category. Using the two-way contingency table of a word $\boldsymbol{t}$ and a category c − i) A is the number of times t and c co-occur, ii) B is the number of times t occurs without c, iii) C is the number of times c occurs without t, iv) D is the number of times neither c nor t occurs, and vi) N is the total number of sentences − the word goodness measure is defined as follows:

$$\chi^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1)$$

From equation (1), words that have a χ^2 test score larger than 2.366 which indicates statistical significance at the 0.5 level are selected as features for respective categories.

4.3 Naïve Bayes Myanmar Text Classifier

The method that is used for classifying documents is Naïve Bayes which is based on Bayesian theorem. The basic idea in Naïve Bayes approach is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. In Naïve Bayes classifier, each document is viewed as a collection of words and the order of words is considered irrelevant. Given a document d for classification, the probability of each category c is calculated as follows:

$$P(c|d) = \frac{P(c)\prod_{w\in d} P(w|c)^{n_{wd}}}{P(d)} \quad (2)$$

where n_{wd} is the number of times word w occurs in document d, P (w|c) is the probability of observing word w given class c, P(c)is the prior probability of class c, and P(d)is a constant that makes the probabilities for the different classes sum to one. P(c)is estimated by the proportion of training documents pertaining to class c and P(w|c) is estimated as:

$$P(w|c) = \frac{1+\sum_{d\in D_c} n_{wd}}{k+\sum_{w'}\sum_{d\in D_c} n_{w'd}} \quad (3)$$

where D_c is the collection of all training documents in class c, and k is the size of the vocabulary (i.e. the number of distinct feature words in all training documents). The additional one in the numerator is the so-called Laplace correction, and corresponds to initializing each word count to one instead of zero. It requires the addition of k in the denominator to obtain a probability distribution that sums to one. This kind of correction is necessary because of the zero-frequency problem: a single word in test document d that does not occur in any training document pertaining to a particular category c will otherwise render P (c|d) zero.

4.4 Execution of Naïve Bayes Myanmar News Classification

We give an example of the execution of our system and we try to classify the input text**: For example:**

Input
Text(T)=>ရွေးကောက်ပွဲဥပဒေအရမညီညွတ်သော သာအမတ်ကိုလွှတ်တော်မှပယ်ဖျက်နိုင်သည်။

1) Preprocessing

Segmented Result=> ရွေးကောက်ပွဲ_ဥပဒေ_အရ_မ_ညီညွတ်_သော_အမတ်_ကို_လွှတ်တော်_မှ_ပယ်ဖျက်_နိုင်_သည်_။

Remove Stopwords=>ရွေးကောက်ပွဲ_ဥပဒေ_မ_ညီညွတ်_အမတ်_ လွှတ်တော်_ ပယ်ဖျက်_

2) Feature Extraction with Chi Square (Example)

Number of example training documents=12

Number of categories=4

ရွေးကောက်ပွဲ{Sport=1, Politic=2}

According to Equation (1),
A=the number of documents that have the feature and belong on the specific category
B=the total number of documents that do not have the particular feature But they belong to the specific category
C=the number of documents that have feature and don't belong to the specific category
D=the number of documents that don't have feature and don't belong to the specific category

Table 1 Sample Features from All Documents

Feature	Category	A	B	C	D	χ^2 Score
ရွေးကောက်ပွဲ	Sport	1	2	2	7	0.14<2.366
	Politic	2	1	8	1	3.7>2.366
ဥပဒေ	Politic	2	1	0	9	7.2>2.366
လွှတ်တော်	Politic	3	0	0	9	3.2>2.3667

We selected a χ^2 Score greater than 2.366 as features.

3) Naïve Bayes Text Classifier
Number of Feature=45
Number of training observation=12
Category=4
Prior Probability

P(Politic)=3/12=1/4, P(Business)=3/12=1/4, P(Entertainment)=1/4, P(Sport)=1/4

Posterior Probability of Feature

P(ရွေးကောက်ပွဲ|Politic)=0.0517241379

P(ရွေးကောက်ပွဲ|Sport)=0.0357142857

P(ရွေးကောက်ပွဲ|Business)=0.0172413793

P(ရွေးကောက်ပွဲ|Entertainment)=0.0194915254

P(ဥပဒေ|Politic)=0.0517241379

P(ဥပဒေ |Sport)=0.0178571428

P(ဥပဒေ|Business)=0.0172413793

P(ဥပဒေ |Entertainment)=0.0169491525

P(လွှတ်တော်|Politic)=0.0344827586

P(လွှတ်တော်|Sport)=0.0178571428

P(လွှတ်တော်|Business)=0.0124137931

P(လွှတ်တော်|Entertainment)=0.0169491525

P(T|Politic)=1/4*0.05*0.05*0.03=0.000023
P(T|Business)=1/4*0.01*0.01*0.01=0.000003
P(T|Entertainment)=1/4*0.01*0.01*0.01=0.000002
P(T|Sport)=1/4*0.03*0.01*0.01=0.0000028

P(T|Politic) is the biggest. Therefore, this document is classified as politics.

5. KNN Myanmar Text Classification Process

K nearest neighbors is one of the statistical learning algorithms that have been used for text classification [12, 13]. KNN is known as one of the top classification algorithms for most language. The k-nearest-neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples.

At the beginning of the classification, it is necessary to select the document that will be carried out for classification and include it in the belong category [9]. For the selected document, its weight value also must be determined by TF-IDF method with example Table 2, as well as for all other documents. TF-IDF is calculated as

$$a_{td} = tfidf(t,d,D) = tf(t,d).idf(t,D) \quad (4)$$

$$tf(t,d) = \frac{f_{t,d}}{\max\{f_{t',d}:t'\epsilon d\}} \quad (5)$$

$$idf(t,D) = log_2\left(\frac{N}{df_t}\right) \quad (6)$$

where a_{td} = is the weight of term t in document d
N=total number of documents in the corpus N=|D|
$f_{t,d}$=frequency of term t in document d
df_t=the document frequency of term t in the collection

Table 2 Sample Term Frequency for Three Documents

Term	Doc1	Doc2	Doc3
ရွေးကောက်ပွဲ	1	1	3
ဥပဒေ	1	0	1
လွှတ်တော်	2	0	1
Category	Politic	Sport	Politic

Classification process writes data of the selected document in a weight matrix to its end. Writing down all data in the same weight matrix has resulted with optimization and reducing the total time of calculation.

In the next step of the process it is necessary to determine K value. K value of the KNN algorithm is a factor which indicates a required number of documents from the collection which is closest to the selected document. If $K = 1$, then the object is simply assigned to the class of that single nearest neighbor. The classification process determinates the vectors distance between the documents by using the following equation [11]:

$$d(x,y) = \sqrt{\sum_{r=1}^{N}(a_{rx} - a_{ry})^2} \quad (7)$$

where d(x,y) is the distance between two documents, N is the number unique words in the documents collection, a_{rx} is a weight of the term r in document x, a_{ry} is a weight of the term r in document y. Pseudo code shows implementation

Pseudo code:

```
for(i=0 i<numberOfDocuments i++)
for(r=0 r<numberOfUniqueWords i++)
d[i]+=(A[r,i]-A[r,( numberOfDocument-1)])^2
end
        d[i] = Sqrt( d[i] )
end
```

Smaller Euclidean distance between the documents indicates their higher similarity. Distance 0 means that the documents are complete equal.

7. Experimental Work

7.1 Data used for Experiment

The experiment is conducted using data collected from Myanmar news websites which contain news for all pre-defined categories. The proposed system is relied on supervised learning. The training set consists of over 1200 news and test set contains 150 news for each category. Both training and test data include Myanmar news which is composed of pure text data and speech transcriptions. An average number of sentences per document are 10.

Table 3 Data Set

Data	Politic	Business	Entertainment	Sport
No: Training Doc	300	300	300	300
No: Testing Doc	150	150	150	150

7.2 Performance Measures and Result

In this paper, a document is assigned to only one category. Precision, recall and F-measure are used as performance measures for the test set. In measuring performance, precision, recall and Fmeasure for each category are calculated as the following equations.

$$\text{Precision (P)} = \frac{a}{b} \quad (8)$$

$$Recall\ (R) = \frac{a}{c} \quad (9)$$

$$Fmeasure(F1) = \frac{2PR}{P+R} \quad (10)$$

a=no. of correctly classified documents to a category

b=Total no. of documents labeled by the system as that category

c= no. of documents of that category in training data

Table 4 Experimental Results for Two Classifiers

Category	P(%)		R(%)		F1(%)	
	NB	KNN	NB	KNN	NB	KNN
Politic	88	91	89	89	88	89
Business	78	78	87	73	82	75
Entertainment	79	81	71	79	74	79
Sport	81	91	83	84	81	87

KNN classifier is found to be most vulnerable with respect to number of features and feature selection method. Naive Bayes is observed to be second top performing classifier. However, its performance also depends on the choice of feature selection method and number of features.

The failure in classification process is caused by the amount of training corpus, and the problem of segmentation. In experiment, our results demonstrate that the combination of KNN algorithm and TF-IDF method can work well on Myanmar News data set.

8. Conclusion

This paper has described a text classification method for Myanmar language. In the proposed system, pre-classified training documents are manually collected for each category and used them as training data. And then, it classifies test documents. This could be a significant method in text learning because this can reduce the cost and time for classifying documents by hand. Therefore, this method can be used in areas where low-cost text classification is required.

The classification performance K-NN classifier is far better than naïve basian classifier when learning parameters and number of samples are small. Almost all the important techniques for classification such as decision trees, Bayes methods, nearest neighbor classifiers, and SVM classifiers have been extended to the case of text data. Research shows that KNN has the maximum accuracy as compared to other classification methods. But problem of KNN is that its time complexity is high but better accuracy than others.

In the future, more and more training data are going to be trained. The accuracy will be higher. We have to test other category and other classifier such as SVM, neural network, Rocchio etc.

9. References

[1] Youngjoong Ko and Jungyun Seo, "Automatic Text Categorization by Unsupervised Learning".

[2] S M Kamruzzaman and Chowdhury Mofizur Rahman, "Text Categorization using Association Rule and Naïve Bayes Classifier".

[3] Eibe Frank and Remco R. Bouckaert, "Naïve Bayes for Text Classification with Unbalanced Classes".

[4] Chee-Hong Chan, Aixin Sun and Ee-Peng Lim, "Automated Online News Classification with Personalization", December 2001.

[5] M. IKONOMAKIS, S. KOTSIANTIS and V. TAMPAKAS, "Text Classification Using Machine Learning Techniques", August 2005

[6] S.Niharika , V.Sneha Latha and D.R.Lavanya, "A Survey on Text Categorization", 2012.

[7] Hiroshi Shimodaira, "Text Classification Using Naive Bayes", February 2015.

[8] Anuradha Purohit, Deepika Atre, Payal Jaswani and Priyanshi Asawara, "Text Classification in Data Mining", June 2015.

[9] Ahmed Faraz, "An Elaboration of Text Categorization And Automatic Text Classification Through Mathematical and Graphical Modeling", June 2015.

[10] Bruno Trstenjak, Sasa Mikac, Dzenana Donko, "KNN with TF-IDF Based Framework for Text Categorization", 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

[11]K. Mikawa, T. Ishidat, M.Goto, "A Proposal of Extended Cosine Measure for Distance Metric Learning in Text Classification", 2011.

[12] Y. Yang and X. Liu, "A re-examination of text categorization methods". In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),42-49, 1999

[13] Riyad Al-Shalabi, Ghassan Kanaan, Manaf H. Gharaibeh, "Arabic Text Categorization Using kNN Algorithm", The International Arab Journal of Information Technology, Vol 4, P 5-7,2015.

[14] Meenakshi and Swati Singla, "Review Paper on Text Categorization Techniques", SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – EFES April 2015