Automatic Myanmar News Classification Using Naïve Bayes Classifier

Aye Hninn Khine
ayehninnkhine@ucsy.edu.mm

Dr. Khin Thandar Nwet
khinthandarnwet@ucsy.edu.mm

University of Computer Studies, Yangon

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. In general, text classification plays an important role in information extraction, text summarization, information retrieval, and question and answering. The goal of text classification is to classify documents into a certain number of pre-defined categories. This paper illustrates the text classification process using machine learning technique and it aims to express how to automatically classify Myanmar news using Naïve Bayes algorithm. The proposed system is an initial attempt of the text classification system for Myanmar language. News corpus is used for training and testing purpose of the classifier. Chi square feature selection algorithm is used in the proposed system to select the most relevant features from training data.

Keywords: Text Mining, Text Classification, Natural Language Processing, Machine Learning

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. So, 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.

The remaining parts of this paper are organized as follows. In section 2, the nature of Myanmar language is discussed. Theory background is described in section 3 and the related works are explained in section 4. Then, the overview of the proposed system is shown in section 5 and the proposed algorithm is illustrated in section 6. In the later sections, section 7 and

section 8, the experimental work is described and the paper is concluded specifically.

2. Text Mining and Automatic Text Classification

Text mining refers to the discovery of previously unknown knowledge that can be found in text collections. It is also a form of data mining but it extracts patterns from natural language text rather than databases. It is an interdisciplinary field that borrows techniques from the general field of data mining and it additionally combines methodologies from various areas such as Information Extraction (IE) and Information Retrieval (IR).

Text mining can be defined in three steps theoretically. These three steps are implemented by text mining researchers as follows:
(a)Analysis of larger quantities of text
(b)Detection of usage patterns from text
(c)Extraction of useful and correct information from detected usage patterns

According to this definition, text classification is an instance of text mining and it intends to organize a set of texts, identify the structure of a text collection and group text documents according to their common features. In this way, unstructured repositories obtain some structure, the labeling, search and browsing of documents is enabled and the text data analysis becomes efficient and effective.

Intuitively text classification is the task of classifying a document under a pre-defined category. More formally, if d_i is a document of the entire set of documents D and $\{ c_1, c_2, ..., c_n \}$ is the set of all the categories, then text classification assigns a document $\mathbf{d_i}$ to one category $\mathbf{c_j}$. The basic phases in text classification include preprocessing the features, extracting the relevant features against the features in a training corpus, and finally categorizing a set of documents into predefined categories. A major characteristic of text classification is high dimensionality of the feature space.

Automatic text classification can be done by two approaches, namely, supervised learning and unsupervised learning. In the supervised learning, training data set is needed. On the other hand, unsupervised learning allows approaching the problems with little or no idea what the results should look like. The structure from data can be derived by clustering the data based on relationships among the variables in the data. In the proposed system supervised learning method is used.

3. 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. Myanmar Word Segmentation Version 1.0 developed by UCSY NLP Lab is applied to segment the sentences in the proposed system.

4. Related Work

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 M 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, Chee-Hong Chan , Aixin Sun and Ee-Peng Lim presented classifier to classify news articles from the Channel News Asia[3]. 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, TfxIDf algorithm is used in this classifier. Support Vector Machine is applied in the classification stage of this classifier.

5. Overview of The Proposed System

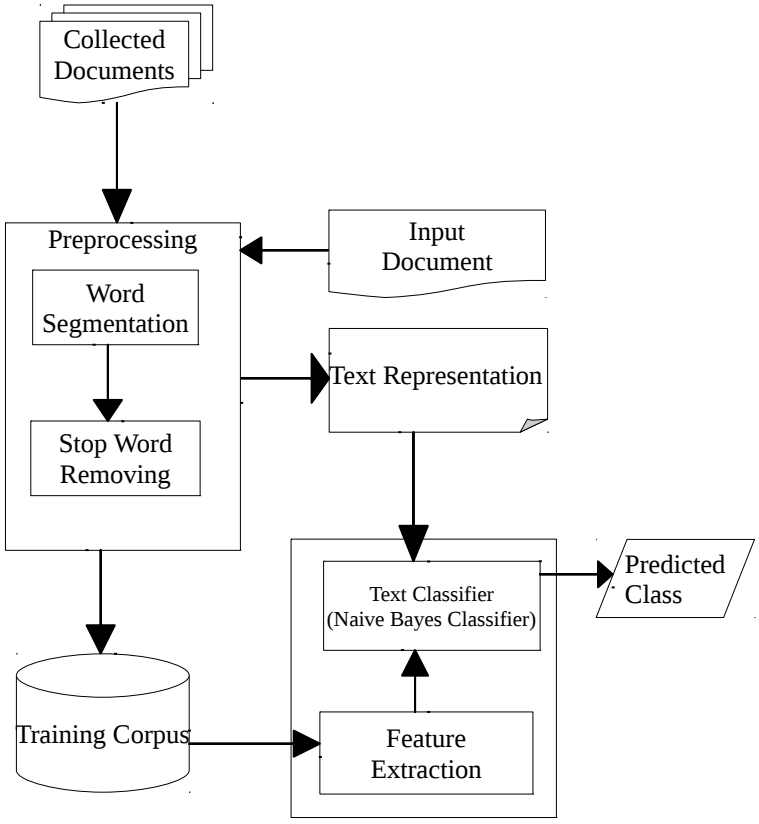

Fig 1: Proposed System Design

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. So, news from Myanmar media websites such as news-eleven.com, 7daydaily.com and popularmyanmar.com are manually collected for each category.

5.1 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.

First, sentences in collected documents are segmented into words using UCSY Word Segmentation Version 1.0. And then, stop words, punctuations and special characters are removed from the collected documents.

Example:

Input=>ရွေးကောက်ပွဲဥပဒေအရ
မည်ညွတ်သောအမတ်ကိုလွှတ်တော်မှပယ်ဖျက်နိုင်သည်။

Segmented Result=>ရွေးကောက်ပွဲ_ဥပဒေ_အရ_
မ_ည်ညွတ်_သော_အမတ်_ကို_လွှတ်တော်_မှ_
ပယ်ဖျက်_နိုင်_သည်_ ။

Remove Stopwords=>ရွေးကောက်ပွဲ_ ဥပဒေ_မ_
ည်ညွတ်__အမတ်__လွှတ်တော်__ပယ်ဖျက်_

5.2 Feature Selection and Text Classifier

5.2.1 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(X^2) statistic with respect to the category. Using the two-way contingency table of a word ***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:

$$X^2(t,c)= \frac{Nx(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

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

5.2.2 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.

6. Proposed Algorithm

The proposed algorithm of the automatic Myanmar news classification system is illustrated in figure 2.

Step1 : Preprocessing

-Segment input text

-Remove stop words, punctuations and special characters from input text and create text represenation

Step2: Feature Extraction

-extract features from training corpus

for each category c in all categories

for each word w in training corpus

$$X^2(w,c)= \frac{Nx(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

if $X^2(w,c)>2.366$

store this word as "feature" for respective category

end for

end for

Step3: Calculate Prior Probability

-calculate prior probability of each category

for each category c in all categories

$$P(c)= \frac{Number\ of\ training\ documents\ whose\ category\ is\ c}{Number\ of\ total\ training\ documents}$$

end for

Step4: Calculate Posterior Probability

-calculate posterior probability of each feature for each category from input text

for each category c in all categories

for each feature word w in all features

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

end for

end for

Step5: Calculate Probability of Each Category

-calculate probability of each category for input text

for each category c in all categories

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

end for

Step6: Choose correct category for input text

c'=arg max score(c_i)

Figure 2: Proposed Algorithm for Automatic Myanmar News Classification

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 600 news and test set contains 100 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 is 10.

	Politic	Business	Entertainment	Sport
No: Training Doc	150	150	150	150
No: Testing Doc	100	100	100	100

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 performace measures for the test set. In measuring performance, precision, recall and F-measure for each category are calculated. For the text classification process, precision of a category for test set is the ratio of the number of correctly classified documents to that category to the total number of documents labeled by the system as that category. Recall is the ratio of the number of

correctly classified documents to the number of documents of that category in training data. The F1 score can be interpreted as a weighted average of the precision and recall. They are calculated by utilizing the following equations:

Precision:

$$\frac{\text{Number of correctly classified documents to a category}}{\text{Total number of documents labeled by the system as that category}} \quad (4)$$

Recall:

$$\frac{\text{Number of correctly classified documents to a category}}{\text{the number of documents of that category in training data}} \quad (5)$$

F_1(recall,precision):

$$\frac{2\,x\,(recall\,x\,precision)}{recall + precision} \quad (6)$$

The experimental result of the test data set is shown in the following figure.

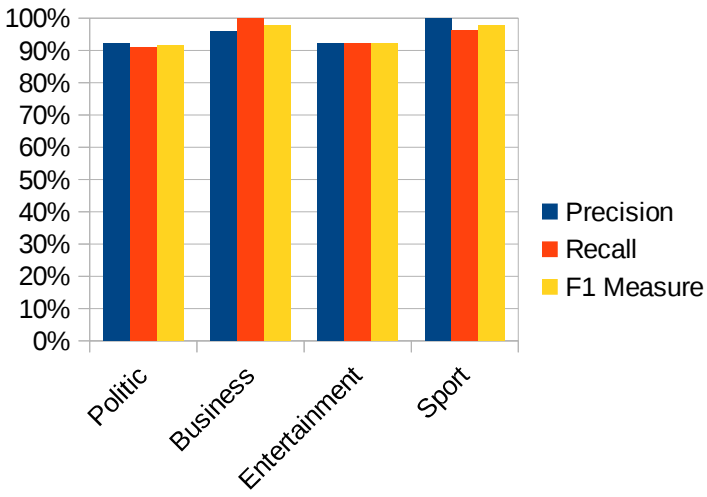

Figure 3: Experimental result on test data set

The experiment shows that classification process by using the proposed method from the mentioned corpus, received about 92% overall accuracy in classifying Myanmar news. The 8% failure in classification process is caused by the amount of training corpus, and the problem of segmentation.

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.

Since the proposed system uses supervised approach, it strongly depends on the size of training data. According to the result, it works well only on well-defined categories and not bounded category is needed to add for invalid test data.

To improve the performance of the proposed system, adding and training more categories will be carried out and more data will be added to the training set. Hence, our proposed system of Myanmar news classification can be considered to be useful and applicable for other research efforts in natural language processing.

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, “Naive 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