MACHINE LEARNING ALGORITHMS FOR MYANMAR NEWS CLASSIFICATION

Khin Thandar Nwet

Department of Computer Science, University of Information Technology, Yangon, Myanmar

ABSTRACT

Text classification is a very important research area in machine learning. Artificial Intelligence is reshaping text classification techniques to better acquire knowledge. In spite of the growth and spread of AI in text mining research for various languages such as English, Japanese, Chinese, etc., its role with respect to Myanmar text is not well understood yet. The aim of this paper is comparative study of machine learning algorithms such as Naïve Bayes (NB), k-nearest neighbours (KNN), support vector machine (SVM) algorithms for Myanmar Language News classification. There is no comparative study of machine learning algorithms in Myanmar News. The news is classified into one of four categories (political, Business, Entertainment and Sport). Dataset is collected from 12,000 documents belongs to 4 categories. Well-known algorithms are applied on collected Myanmar language News dataset from websites. 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 method, chi square algorithm achieves comparable performance across a number of classifiers. In this paper, the experimental results also show support vector machine is better accuracy to other classification algorithms employed in this research. Due to Myanmar Language is complex, it is more important to study and understand the nature of data before proceeding into mining.

KEYWORDS

Text Classification, Machine Learning, Feature Extraction

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 labelled documents. For the training data set, news from Myanmar media websites are manually collected, labelled and stored in the training data set. Chi Square function is used as a feature selection method and classification algorithm is applied in implementing the text classifier. In Naïve Byes text classifier, word frequencies are used as features.

This paper used k-Nearest Neighbors, Naïve Bayes and SVM Classifier in Myanmar text classification. 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. The related works are explained in section 2. In section 3, the nature of Myanmar language is discussed. Then, the overview of the proposed system is shown in section 4 and machine learning algorithms for text classification is illustrated in section 5. In the later sections, section 6 and section 7, the experimental work is described and the paper is concluded specifically.

2. RELATED WORK

Many algorithms have been applied for Automatic Text Categorization. Most studies have been devoted to English and other Latin languages. However, there is no research in comparative study of machine learning for Myanmar text.

The paper written by K. T. D. Nwet, A.H.Khine and K.M.Soe, naïve bayes and KNN are used for automatic Myanmar News classification in Myanmar Language. KNN is better accuracy for text classification. A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification 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].
This paper used k-Nearest Neighbors, Naïve Bayes and SVM Classifier in Myanmar text classification. Its aims at making a comparative study between mentioned algorithms on a Myanmar data set. It also shows support vector machine is better accuracy to other classification algorithms employed in this research.

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. These language is low resource language. 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. In this paper, Syllable level Longest Matching approach is used for Myanmar word segmentation. Due to Myanmar Language is complex language, it is more important to study and understand the nature of data before proceeding into mining.

4. OVERVIEW OF THE PROPOSED SYSTEM

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 news 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.

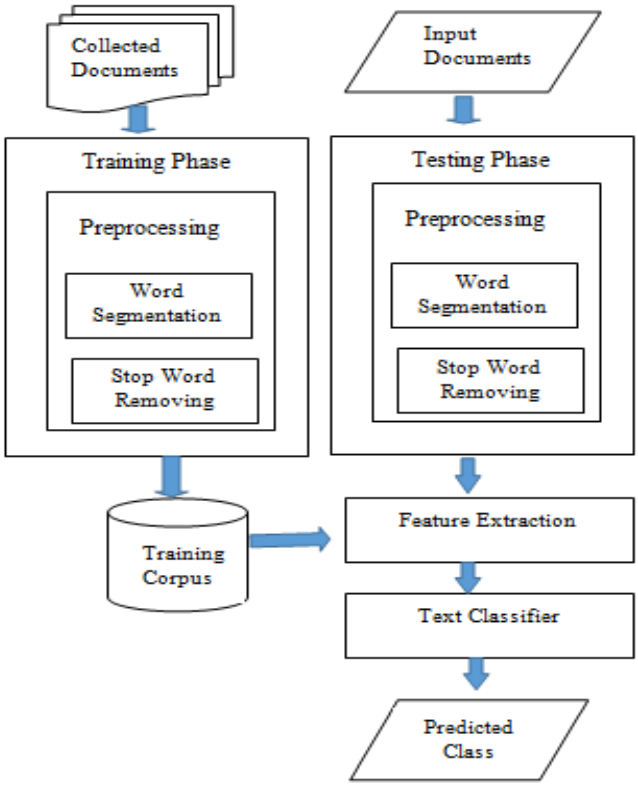

Figure 1 Proposed System Design

5. MACHINE LEARNING APPROACH

5.1. 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 (χ^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:

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

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.

5.2. Naïve Bayes

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, 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.

5.3. K-Nearest Neighbour

K nearest neighbors(KNN) 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 Chi Square method.

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]:

$$d(x,y) = \sqrt{\sum_{r=1}^{N}(a_{rx} - a_{ry})^2} \quad (4)$$

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.

5.4. Support Vector Machine

The Support Vector Machine (SVM) algorithm is one of the supervised machine learning algorithms that is employed for various classification problems [16]. It has its applications in credit risk analysis, medical diagnosis, text categorization, and information extraction. SVMs are particularly suitable for high dimensional data. There are so many reasons supporting this claim. Specifically, the complexity of the classifiers depends on the number of support vectors instead of data dimensions, they produce the same hyper plane for repeated training sets, and they have better generalization. SVMs also perform with the same accuracy even when the data is sparse. SVMs tend to use over-fitting protection mechanism that is independent of the dimensionality of the feature space.

6. EXPERIMENTAL WORK

6.1. Data

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 12,000 news and test set contains 500 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. These experiments were carried out using an open source Python library, 'scikit-learn' [17].

Table 1. Data

Data	Politic	Business	Entertainment	Sport
No: Training Doc	3,000	3,000	3,000	3,000
No: Testing Doc	500	500	500	500

6.2. Performance

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.

$$\text{Precision (P)} = \frac{a}{b} \quad (5)$$

$$Recall\ (R) = \frac{a}{c} \quad (6)$$

$$Fmeasure(F1) = \frac{2PR}{P+R} \quad (7)$$

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 2. Experimental Results for Three Classifiers

Category	F1(%)		
	NB	KNN	SVM
Politic	84	86	88
Business	78	75	85
Entertainment	79	79	82
Sport	81	85	87

SVM classifier is found to be most vulnerable with respect to number of features and feature selection method. KNN 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, the results also show support vector machine is better accuracy to other classification algorithms employed in this research. This algorithm gives high accuracy, nice theoretical guarantees regarding overfitting. Especially, it is popular in text classification problems where very high-dimensional spaces are the norm. It therefore seems natural that SVM tends to perform better than Naïve Bayes and KNN for this task.

7. CONCLUSIONS

Support vector machine algorithm is better to other classification algorithms employed in this research. Due to Myanmar Language is complex language, it is more important to study and understand the nature of data before proceeding into mining. With the increasing amount of news data and need for accuracy, the automation of text classification process is required. Another interesting research opportunity bases on Myanmar language news classification model with deep learning systems. Deep learning algorithms such as GloVe or Word2Vec are also used in order to obtain better vector representations for words and improve the accuracy of classifiers trained with traditional machine learning algorithms. In the future, deep learning algorithms can be used to test on the same dataset for better results.

REFERENCES

[1] Y. Ko and J. Seo, "Automatic Text Categorization by Unsupervised Learning", 2000.

[2] S M Kamruzzaman and C. Mofizur Rahman, "Text Categorization using Association Rule and Naïve Bayes Classifier",2001

.

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

[4] C. Chan, A. 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] H. 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] B. Trstenjak, S. Mikac, D. 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 and 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] R. Al-Shalabi, G. Kanaan and 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

[15] K. T. D. Nwet, A.H.Khine and K.M.Soe "Automatic Myanmar News Classification", Fifteenth International Conference On Computer Applications, 2017.

[16] M. Thangaraj and M. Sivakami, "Text classification Techniques: A Literature Review", Jornal of Information, Knowledge and Management, Vol 13, 2018

[17] Fabian Pedregosa, Ga¨el Varoquaux, Alexandre Gramfort, Vincent Michel,Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss,Vincent Dubourg, et al. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12(Oct):2825{2830, 2011.

AUTHORS

The author's research interests are Machine Translation and Speech synthesis. She has been supervising Master thesis on Natural language processing such as Information Retrieval, Morphological Analysis, Summarization, Parsing, and Machine Translation.