Myanmar Text Classifier Using Genetic Algorithm

Thit Thit Zaw, Khin Mar Soe
University of Computer Studies, Yangon
sophie.thitzaw@ucsy.edu.mm, khinmarsoe@ucsy.edu.mm

Abstract

Text Classification is the task of automatically assigning a set of documents into certain categories (class or topics) from a predefined set. This also play important role in natural language processing and also crossroad between information retrieval and machine Learning. The dramatic growth of text document in digital form news website make the task of text classification more popular over last ten year. The application of this method can be found in spam filtering, question and answering, language identification. This paper presents the idea of text classification process in term of using machine learning technique and illustrates how Myanmar news documents were classified by applying genetic algorithm. The applied system will be used Myanmar online news articles from Myanmar news website for the purpose of training and testing the system. Term frequency inverse document frequency (tf_idf) algorithm was used to select related feature according to their labelled document which is also applied in many text mining methods.

Keywords: *Text Mining, Text Classification, Natural Language Processing, Machine Learning, Genetic Algorithm*

1. Introduction

Over the recent year, as the amount of online text documents raised dramatically, the need to access them in more flexible ways was also increased. Due to this, text mining becomes one of the important key factors to satisfy the need to manage or extracting suitable information from text data. The process can be defined as text mining process and which include text classification, text summarization, text clustering, and sentiment analysis.

Text Classification technique has been applied to text filtering, news articles categorization, document classification, language filtering and spam filtering and etc. These systems are Language oriented and English, Arabic, French and other European and Asian Languages were well developed .However, text classification for Myanmar Language has still challenging task due to some difficulties in tokenizing, stemming, feature selection of Myanmar words.

In this paper, automatic classification of Myanmar new articles were present. The proposed system will use semi-supervised Learning approach will be applied. The system will use Genetic algorithm to implement the classification of news articles according to their content into already predefined categories such as Politics, Business, Sport, and Entertainment. Myanmar news articles were collected from Myanmar news local websites and tokenize, stemming and stop words collection were done manually. Term Frequency Inverse Document Frequency algorithm (tf_idf) will be applied in preprocessing stage where feature that are related to each predefined categories selected.

The remaining parts of the paper are organized as follow. In section 2, Myanmar Language nature will be discussed and theory background can be seen in section 3 .The related work and overview of the proposed system will be explained in section 4 and 5 respectively. Section 6 contain the information about the proposed algorithm .For last two sections , section 7 and section 8 , experimental work is described and the paper will be concluded .

2. Text Mining and Automatic text classification

Text mining which can also know as knowledge discovery of text data is the process of extracting the useful and interesting information from unstructured text documents and is also a subset of the larger fields of data mining. Text mining is a multidisciplinary field, where it involves the technique from other areas like information retrieval, text analysis, information extraction, clustering, classification, categorization, visual-ization, machine learning, and data mining.

Text classification is the instance of the text mining. Text classification process includes preprocessing the documents, extracting relevant feature according to the feature in the training corpus and applied classification algorithm to classify document into predefined categories. Basically , text classification classify document d_j from the entire collection of the document D and classify it into one of the category from { $c_1,c_2,c_3 \ldots c_i$ }.

Text classification tasks can be divided into three sorts:

1).**supervised document classification** where some external mechanism (such as human feedback) provides information on the correct classification for documents,

2).**unsupervised document classification** (also known as document clustering), where the classification must be done entirely without reference to external information

3).**semi-supervised document classification**, where parts of the documents are labeled by the external mechanism.

The proposed system applied semi supervised document classification method to categorize Myanmar Language news from news website into four categories -politics, business, entertainment and sport.

3. Myanmar Language

Myanmar Language is the official Language of Republic of Union of Myanmar .Myanmar alphabet consists of 33 letter and 12 vowels and has circular shape. It was written from left to right and no space between each words .However, modern written style added spaces between each word to increase readability. Although the basic word order of Myanmar language is free but usually follow subject-object-verb order and usually end with verb. The language is monosyllabic and its sentence contains several preposition adjunctions.

As there are no strict rules about spaces in sentences, many people write spaces between letters as they feel fit which make the process of words segmentation of Language make difficult.

4. Related Work

S.M.Khalessizadeh, R.Zaefarian, S.H.Nasseri, and E.Adril ,[2] used genetic algorithm for topic based on Concept Distribution. They use GA in concept weighting in standard text in Persian language. According to experimental result, they compare genetic algorithm with traditional TF-IDF method and it shows that GA has notable improvement

S.M.Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan,[3] presented new algorithm for text classification using data mining and compare with other classification algorithms like Naive Bayes classier, Genetic Algorithm, Decision Tree, Apriori algorithm .The experimental results show that the proposed system work as a successful classifier with 90% accuracy rate. The paper suggest that the proposed system work well with large dataset.

5. Overview of the Proposed System

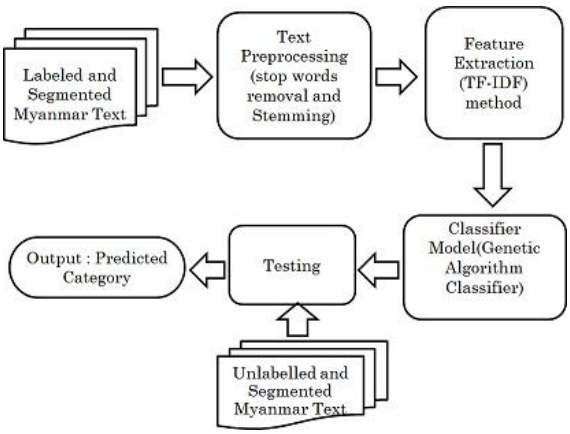

Figure 1. Proposed System Design

According to the figure, the propose system has two phases: training and classification phases. In training phases, the features from corpus that are related to their labeled categories are extracted and store for later use in classification process. Then, in classification phase, unknown text documents were used to classify into predefined category according to their content.

As the propose system is semi-supervise learning classification approach, the collection of the text document were necessary. These text documents were collected from news-eleven.com; 7daydaily.com and popularmyanmar.com are manually collected for each category.

5.1 Preprocessing

Preprocessing in text mining process play important role in text mining process. In preprocessing step, selecting the significant keywords that carry the meaning and discarding the words that do not contribute to distinguishing between the

documents. The main objective is to obtain the key features from online news text documents and to enhance the relevancy between words and documents and the relevancy between words and categories.

First of all, the sentences in preprocessing stage were segmented manually according to segmentation rules. The proposed system take input words which were already segmented and removed stop words which is done manually. The segmentation rules is that words were divided according to noun, verb, Adj and Adv surfix words such as များ, တို့ , တွေ , မှု ,ခြင်း and these segmentation was done by manually. The collected stop words, special character, punctuation are removed from the collection of documents.

Example:

- Input (Segmented Text) => ၂၁ - ရာစု - ပင်လုံ-ညီလာခံ - ကို - ဝ - ကိုယ်စားလှယ် - အဖွဲ့-ကျောခိုင်း - ပြီး - နေရပ် - ပြန်

  Then, the collected stop words like နဲ့, တွေ, ကြောင့် were removed.

- Stop words removal=> ရာစု-ပင်လုံ-ညီလာခံ-ဝ-ကိုယ်စားလှယ်-အဖွဲ့-ကျောခိုင်း-နေရပ်-ပြန်

5.2 Feature Selection and Text Classifier

5.2.1 Feature Selection

In the feature selection process, it is important to collect the words that are relevant to their labeled category are selected as feature .These term or features that were collected during feature selection were used in the classification process. The classification processes rely on the term of the feature of the given category to calculate the term value by indexing the feature from each category.[1]

In Myanmar Text classifier system, the weight and value of each term are calculated by using TF-IDF (term frequency Inverse Document Frequency) algorithm. TF-IDF is generally a content descriptive mechanism for the documents. The term frequency (TF) is the number of times a term (word) occurs in documents. Inverse Document frequency (IDF) measure how important a term is? The concepts of term frequency and inverse document frequency are combined, to produce a composite weight for each term in each document.

TF-IDF were calculated as

TF : (1)

$$\frac{\text{Number of times terms t appear in a document}}{\text{total number of terms in the documents}}$$

IDF : (2)

$$\text{Log}\left(\frac{\text{total number of document}}{\text{Number of documents with term t in it}}\right)$$

$$\boldsymbol{TF - IDF} = TF * IDF \qquad (3)$$

Where, TF is the term Frequency and IDF is the Inverse Document Frequency and TF-IDF is the Term Frequency Inverse Document Frequency.

5.2.2 Text Classifier

The method that is used to classify is Genetic Algorithm. The algorithm is based on Darwin's natural selection processes which inspire the biological evolution process. GA is the basically optimization technique which try to find out what is the best result among the given input. Genetic Algorithm is an algorithm which makes it easy to search a large search space. By implementing this Darwinian selection to the problem, only the best solutions will remain as narrowing the search space [1].

Genetic algorithm for text classification process work as

1. [Start] Generate random population of n chromosomes

2. [Fitness] Evaluate the fitness f(x) of each chromosome x in the population with WTSD (weight Term Standard Derivation) fitness function.

$$\textbf{WTSD}(d_i)=\sum \frac{wd_ic_j(x_{j,k}-\bar{x}.wd_ic_j)^2}{(n-1).\bar{w}.maxfrequent(d_i)} \qquad (4)$$

In Equitation [4], di is document that were classified and wd_ic_j is weight of the term and $x_{j,k}$ is number of frequent that term j accrued n represent total number of terms in a document and $\bar{x}$ is the mean of frequent of the term and $\bar{w}$ is an average of weight of term in document.

3. [New population] Create a new population by repeating following steps until the new population is complete

(1). [Selection] Select two parent chromosomes from a population according to their fitness.
(2). [Crossover] With a crossover probability cross over the parents to form new offspring.
(3). [Mutation] With a mutation probability mutate new offspring at each locus.

4. [Accepting] Place new offspring in the new population for a further run of the algorithm.

5. [Replace] Use new generated population for a further run of the algorithm

6. [Test] If the end condition is satisfied, stop, and return the best solution in current population, otherwise go to step 2.

6. Proposed Algorithm

The purpose algorithm of the Myanmar Text Classifier Using Genetic Algorithm was illustrated in following figure.

Step1: Segmentation
- Segment the input text according to segmentation rules manually.

Step2: Stop word removal
- Stop words were collected during the manual segmentation process were removed from the input text.

Step3: Feature Extraction
- extract features from training corpus for each category c in all categories
- apply tf-idf algorithm to determine weight of each term

Tf = TF * IDF

Step4: Training Classifier
- applying the genetic algorithm classifier to the text document to classify
- Weight Term Standard Deviation was used as fitness function.

Step5: Testing the Classifier
- the classifier were tested with unknown document and training document to classify these into their related category.

Figure 2. Proposed Algorithm for Myanmar Text Classifier using Genetic algorithm

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 training set consists of over 800 news and test set contains 25 news for each category. Both training and test data include Myanmar news which is composed of pure text data and speech transcriptions.

	Politics	Business	Entertainment	Sport
No of training doc	200	200	200	200
No of testing doc	25	25	25	25

7.2 Performance Measures and Result

Performance measure for the test set is measure by using precision, recall, F-measure are use. Our proposes system assigned each document to only one category which is related to their content. So, precision, recall, F-measure were calculated for each category. 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: (5)

$$\frac{Number\ of\ correctly classified\ documents\ \text{to}\ acategory}{Total\ number\ of\ documents labeled\ by the\ system as that\ category}$$

Recall: (6)

$$\frac{Number\ of\ correctly\ classified\ documents\ \text{to}\ acategory}{the\ number\ of\ documents\ of\ that\ category\ \text{in}\ training\ data}$$

F1(recall, precision): $2.\frac{(precision.recall)}{(precision+recall)}$ (7)

The experiment shows that classification process by using Genetic Algorithm for collected

corpus is received about 90% overall accuracy in classifying Myanmar news. The 10% failure in classification process is caused by the amount of training corpus, and the problem of segmentation.

8. Conclusion

Text classification is a vital system in natural language processing and information extraction. As the role of text classification also become important, it became successful in wide variety of real world applications. Feature Selection methods also help to reduce the problem of dimensionality in text classification application. Although text classification is well researched area in text mining area, it still needs many developments in the feature preparation and classification engine to accelerate the classification performance for a specific application. Although the propose system work well with defined categories, document which are not bound to each of the category is not going to work with the system.

The propose system can be improved by adding more data to the training documents that it can increase the accuracy of the classification task.

Related work

[1] Aye Hnin Khine, "Automatic Myanmar news classification using Naïve Bayes Classifier", UCSY, Myanmar, 2016

[2] S.M.Khalessizadeh, R.Zaefarian,S.H.Nasseri and E.Adril, "Genetic Mining: Using genetic algorithm for Topic based on Concept Distribution", International journal of Mathematical Computational, Physical, electrical and Computer Engineering Vol: 2, No:1, 2008.

[3] S.M.Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan, "Text Classification Using Data Mining", ICTM2005.

[4] Mohammed Abbas Kadhum, "Using Genetic Algorithm for calssification of flower plants".

[5] Monica Bali, Deipali Gore, "A Survey on Text Classification with Different Types of Classification Methods", on International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 5, May 2015