MYANMAR TEXT CLASSIFIER USING GENETIC ALGORITHM

THIT THIT ZAW

M.C.Sc. **July 2018**

MYANMAR TEXT CLASSIFIER USING GENETIC ALGORITHM

BY

THIT THIT ZAW

B.C.Sc. (Honours)

Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Computer Science (M.C.Sc.) Of the University of Computer Studies, Yangon

July 2018

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to those who helped me with various aspects of conducting research and writing this thesis.

Firstly, I would like to express my deepest gratitude to **Dr. Mie Mie Thet Thwin**, Rector, the University of Computer Studies, Yangon, for her kind permission to develop this thesis.

I am grateful to my thesis supervisor **Dr. Khin Mar Soe**, Professor, NLP Lab, for her suggestions and encouragement to do this thesis and also for her close supervision, invaluable suggestions, kind guidance and constant encouragement during the course of this work.

I also like to express my thanks to our dean **Dr. Thi Thi Soe Nyunt**, Professor and Head of Faculty of Computer Science, for her kind help in this work.

I would like to extend my deepest gratitude to **Daw Ni Ni San**, Lecturer, English Department, for her advice and language editing.

I heartily appreciate the suggestions and recommendations of the teachers who attended all my seminars.

Sincere thanks are also due to my close friends for their kind help, understanding and cooperation throughout the work. Finally, I am most indebted to my beloved family for their support, patience and constant encouragement throughout my studies.

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 plays an important role in natural language processing and also crossroads between information retrieval and machine learning. The dramatic growth of text document in digital form news website makes 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 book presents the idea of text classification process in term of using machine learning technique and illustrates how Myanmar news documents are classified by applying genetic algorithm. The applied system use 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 documents which are also applied in many text mining methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

Today, the use of information technology has significantly increased in developing over the last few years. However, managing these enormous amounts of textual data become more and more difficult. This makes many information technology and web-based services to search for ways to access them in more convenient way of selecting, filtering and managing the unstoppable growing amount of textual data which access is usually critical. Because of these problems, information retrieval and information extraction become the important role in managing this information which was in unstructured data form. Among them, classification of textual data form documents into their related categories is also important way to improve the performance of information retrieval which makes user to access their desire set of documents in more organized way.

Text Classification can be illustrated as assigning documents into one or more predefined categories according to their contents. Classifying large textual data helps in standardizing the platform, makes search easier and relevant, and improves user experience by simplifying navigation. Instead of manually classifying documents or hand-crafting automatic classification rules, text classification uses machine learning methods to learn automatic classification rules based on human-labeled training documents. Manual classification of documents can result in misunderstanding between individual judgements.

Text Classification has been used in many industries in these days. Classification of books in libraries and segmentation of news article are the good example of text classification. With the help of technology, these tasks can be done in minimal amount of time and can reduce substantially and maintain consistency.

There are many number of machine learning algorithms to apply in the classification problem. The machine learning algorithms are Naive Bayes Classifier, Support Vector Machines, Decision Trees, Boosted Trees, Random Forest, Neural Networks and Nearest Neighbor and Genetic algorithm. Nowadays, applying machine learning algorithms in text classification area become more and more popular and these algorithms are widely used in many large companies like Google, Twitter,

Facebook, Amazon and other technological companies. In this thesis, machine learning algorithm namely Genetic Algorithm is used to classify news documents from Myanmar Local News Website which were written in Myanmar Language into their related categories.

1.1 Objective of Thesis

The objectives of this thesis are as follows:

1) To propose a Myanmar Text Classifier based on Genetic Algorithm
2) To extract the features using Term Frequency
3) To enhance the fitness function in GA by using standard deviation
4) To apply the proposed Classifier in classification of Myanmar news

1.2 Overview of Thesis

The proposed system is implemented as a Text Classifier for Myanmar News using semi-supervised Learning method. The preprocessing stage includes segmenting into words, removing stop words and stemming. The TF-IDF (term Frequency–Inverse Document Frequency) algorithm is used in collecting features and calculating predefined value for each feature will later be used in the classification stage. After that, Genetic algorithm is used to classify news documents into their related categories.

1.3 Organization of Thesis

This thesis is composed of five chapters. Chapter 1 presents the introduction and objectives of thesis. Chapter 2 shows background theory of text mining, text classification and machine learning. Chapter 3 describes the details of Myanmar news classification system. Chapter 4 explains the design and implementation of proposed system and experimental result of the classifier. Chapter 5 presents conclusion including challenges, benefit and limitation of the proposed system.

CHAPTER 2
THEORETICAL BACKGROUND

Background Theory of text mining and text classification are mainly discussed in this chapter and also described detail description of the many text classification technique in this section.

2.1 Text Mining

Text mining is process of seeking or extracting the useful information from the textual data. It applies machine learning tools to extract the useful data from data resources in searching for interesting patterns. Nowadays, the growth of the online data make the need to access them in more convenient way rise as most of the online data are in text data. Text mining process is same as data mining, except, the data mining tools are designed to handle structured data whereas text mining can able to handle unstructured or semi-structured data sets such as emails HTML files and full text documents etc. Text Mining is used for finding the new, previously unidentified information from different written resources. [8]

Many Business organizations familiar with structural data which can be text file with row, column and are well organized. These kinds of data were easily visualized and have been used efficiently. Although these structural data were well organized, it didn't even contain the half of the information that the unstructured one carry. So, the role of text mining becomes more important to analyze the unstructured form and primarily text.

As most of the data can be seen in textual form, the text mining is more popular among the business solution to analyze the enormous textual data in the internal file system and web. Finding the interesting patterns from these data source in manual form will be time consuming and need a lot of human labor to review and analyzing these textual data is impossible.

By estimate, 80% of the organizational data come in the form of textual data. So, many organizations analyze their content by applying text mining technique which is the most powerful way. Text mining typically applies machine learning techniques such as clustering, classification, association rules and predictive

modeling. These techniques uncover meaning and relationships in the textual data. Text mining is used in many areas such as competitive intelligence, life sciences, voice of the customer, media and publishing, legal and tax, law enforcement, sentiment analysis and trend-spotting.

The structured data processing challenges are not there in text data processing, as text data is mainly for reading and not for updating. The problem with text data is that it is unstructured. Various preprocessing techniques are required to transform the text data to numerical data so that powerful structured data analytics techniques can be applied. The broad umbrella term Text Mining consists of several sub tasks such as information retrieval, web mining, document clustering, document classification, information extraction, and natural language processing and concept extraction. [12]

Text mining process can be automated and the results from a text mining model can be consistently derived and applied to solve specific problems.
These techniques can be applied in some programs such as:

- Extract key concepts, patterns and relationships from large volumes of textual content
- Spotting trends in textual content based on subjects
- Summarizing content from documents and gain semantic understanding of the underlying content
- Indexing and searching information for the use in analytics process.

2.2 An Introduction to Text Classification

Text Classification is the process of automatic classifying documents into predefined categories based on their content. Machine Learning algorithm and information technique are applied in this area to solve real world problem. In machine learning, text classification has been considered as a classification problem. 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.

The amount of information available on the internet is growing with unstoppable rate. The demand for asking for tools to analyze or organize these

enormous data resources was also increased. Text Classification plays the important role in many information management tasks. Due to these facts, classifying process need to improve performance but also need to maintain accuracy rate are highly desired.

Automatic classification of text documents is useful in the process of dealing with the massive textual data such as library management system, spam filtering system, classifying news sources and analyzing organizational data as most of data input were in textual data. Many theories and machine learning techniques have been applied in text classification.

The classification of documents can be done by human experts, however, human can't be handling these growing dataset and it is time consuming task and human errors in the classification task are also obstacle in the process and these processes are also expensive. Therefore, automatic text classification is the powerful tools for organizing documents and an essential technology for intelligent information system which gains many attentions in recent years. If the need for text classification increases, the business value of the text classification will increase and will become the important technique in digital world. This technique has been applied in classifying new electronic documents, finding interesting information web and guiding user's search through hypertext. Accuracy and efficiency of text classification model is important. The accuracy of the model depends on the number of training set data and its fitness. With so many classification algorithms, the suitability of an algorithm depends on ease of understanding, ease of model building, computation cost and result accuracy measures. [12]

Text classification is the exciting research area in many information retrieval problems such as filtering, routing or searching for relevant information, benefit from the text classification research.

2.3 Operations of Text Classifications

Text classification involves many different steps and processes. A general text classification process normally involves documents collections, preprocessing, feature selection and classification. One of the most relevant and challenging problems is text classification which involves trying to organize text documents into various categories based on inherent properties or attributes of each

text document. This is used in various domains, including email spam identification and news categorization. The concept may seem simple, and if there are only few documents. It can be done just by looking at the documents and gain some idea about what it is trying to indicate. Based on this knowledge, similar documents can be grouped into categories or classes. It's more challenging when the number of text documents to be classified increases to several hundred thousands or millions. This is where techniques like feature extraction and supervised or unsupervised machine learning come in handy. Document classification is a generic problem not limited to text alone but also can be extended for other items like music, images, video, and other media.

To formalize the problem more clearly, a set of classes or categories and several text documents that are basically sentences or paragraphs of text, which form a corpus, are given for the purpose of training. The task is to determine which class or classes each document belongs to. This entire process involves several steps which will be discussing in detail later. Briefly, for a supervised classification problem, some labelled data were needed that could use for training a text classification model. This data would essentially be curated documents that are already assigned to some specific class or category beforehand. It is essential to extract features and attributes from each document and make classification model to learn these attributes corresponding to each particular document and its class/category by feeding it to a supervised machine learning algorithm. Of course, the data would need to be pre-processed and normalized before building the model. Once done, the same process of normalization and feature extraction also need again and then feed it to the model to predict the class or category for new documents. However, for an unsupervised classification problem, it won't essentially need any pre-labelled training documents. The usage of techniques like clustering and document similarity measures to cluster documents together based on their inherent properties and assign labels to them.

There are various types of text classification. This chapter focuses on two major types, which are based on the type of content that makes up the documents:

- Content-based classification
- Request-oriented classification

Both types are more like different philosophies or ideals behind approaches to classifying text documents rather than specific technical algorithms or processes.

Content-based classification is the type of text classification where priorities or weights are given to specific subjects or topics in the text content that would help determine the class of the document. A conceptual example would be that a book with more than 30 percent of its content about food preparations can be classified under cooking/recipes.

Request-based classification is classification that is targeted towards specific user groups and audiences. This type of classification is classification in which the anticipated request from users is influencing how documents are being classified.

2.3.1 Text Data Preprocessing:

All tasks of text mining need numerical representation of the text document. Thus text mining involves additional preprocessing steps to transform the unstructured data to structured data or converting the text format to numerical format first so that data analytics techniques can be applied to them. Heterogeneous sources of text data can be noisy and needs cleansing first.it consist of different tasks which are stop words removal, word segmentation and feature selection.

1) Stop Words Removal

The process of removing words that are commonly occur in the documents corpus of given language is called stop words removal and commonly occur words can be considered as stop words. Stop words are words that need to be filtered out during the task of information retrieval or other natural language tasks. It is the important task in the text preprocessing because excluding some extremely common words which would appear to be of little value in helping select documents matching a user need can help in improving the performance of the text preprocessing process. Words with high frequency which appear in most documents are not helpful for classification either.

2) Word Segmentation

Words segmentation is the process of segmenting sentences into valid words. This process is important in text classification process because the readability of the classifier strongly depend on it. Words segmentation task for Asian language is challenging task because most of the Asian language do not have the definite word boundary like western language .Word segmentation can be done by manually or statistically for those languages. Many researchers have been carried out Myanmar word segmentation in many different ways.

3) Feature Selection

In text classification, the feature selection is the process of selecting a specific subset of the terms of the training set and using only them in the classification algorithm. The goal of feature selection method is to reduce the dimension of data which can gain benefit in making the training faster and it can improve accuracy by removing noisy features. The feature selection process takes place before the training of the classifier. Feature Selection is important either for improving accuracy or for reducing the complexity of the final classifier. The popular feature selection algorithms are Gini Index, Information Gain, Manual Information, Chi-Square and TF-IDF are described in this section.

a) Gini Index

Gini index is the correlation based criterion which attempts to estimate a feature's ability to distinguish between classes. This feature selection method examines the decrease of impurity when using a chosen feature. In this context impurity relates to the ability of a feature to distinguish between the possible classes. The Gini index has low computational requirements, thus making it attractive in high-dimensional data analysis.

b) Information Gain

Another commonly used measure for text feature selection is that of information gain or entropy. Let P_i be the global probability of class I and p_i (w) be the probability of class I, given that the document contains the word w. Let F (w) be

the fraction of the documents containing the word w. Let F (w) be the fraction of the documents containing the word w. The information gain measure i(w) for a given word w is defined as follows:

$$\text{i(w)} = -\sum_{i=1}^{k}.p_i.\log(p_i) + F(w)\sum_{i=1}^{k}P_i(w).\log\big(p_i(w)\big) + \big(1-F(w)\big).\sum_{i=1}^{k}\big(1-p_i(w)\big).\log\big(1-p_i(w)\big) \quad \text{Eq (2.1)}$$

c) Mutual Information

Mutual information (MI) between a term t and a class c is defined by

$$\text{MI (t, c)} = \log\frac{\Pr(t,c)}{\Pr(t),\Pr(c)} \quad \text{Eq (2.2)}$$

To measure the global goodness of a term in feature selection, the users need to combine the category specific scores as

$$\text{MI}_{\max}(t) = \max_i \text{M I}(t,c_i). \quad \text{Eq (2.3)}$$

Alternatively, in some studies, it is also define as

$$\text{MI}_{\max}(t) = \sum_i \Pr(c_i)\text{M I}(t,ci) \quad \text{Eq (2.4)}$$

[17]

d) Chi-Square

X_2 statistics is a different way to compute the lack of independence between the word w and a particular class i. Let n be the total number of documents in the collection ,$p_i(w)$ be the conditional probability of class I for documents which contain w, P_i be the global fraction of documents containing the class I , and F (w) be the global fraction of documents which contain the word w. The X^2 statistics of the word between word w and class i is defined as follows:

$$X_i^2(w) = \frac{n.f(w2).(Pi(w)-pi)^2}{F(w).\big(1-F(w)\big).Pi.(1-Pi)} \quad \text{Eq (2.5)}$$

In the case of mutual information, a global X^2-statistics can be computed from the class-specific values. The average of maximum values can be used in order to create the composite value.

It is noted that X_2-statistics and mutual information are different ways of measuring the correlation between terms and categories. One major advantage of X_2 –statistics over the mutual information measure is that it is a normalized value, and therefore values are more comparable across terms in the same category [17]

e) TF-IDF

TF-IDF short term for term frequency –inverse document frequency is generally a content descriptive mechanism for the documents. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval, text mining, and user modeling. The TF-IDF value increases proportionally t the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

The term frequency will increase if the frequency of the word increases in the corpus. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. One of the simplest ranking functions is computed by summing the TF-IDF for each query term; many more sophisticated ranking functions are variants of this simple model.

TF-IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

The term frequency (TF) is the number of times a term (word) occurs in a document.

$$\boldsymbol{TF} = \frac{\text{Number of times terms t appear in a document}}{\text{total number of terms in the documents}} \qquad \text{Eq (2.6)}$$

Inverse Document frequency (IDF) measure how important a term is for the corpus.

$$\boldsymbol{IDF} = \log\left(\frac{\text{total number of document}}{\text{total number of document}}\right) \qquad \text{Eq (2.7)}$$

The concepts of term frequency and inverse document frequency are combined to produce a composite weight for each term in each document.

$$\boldsymbol{TF - IDF} = TF * IDF \qquad \text{Eq (2.8)}$$

2.3.2 Text Classification Approach

In the second half of the twentieth century, machine learning evolved as a subfield of artificial intelligence that involved the development of self-learning algorithms to gain knowledge from that data in order to make predictions. Instead of requiring humans to manually derive rules and build models from analyzing large

amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models, and make data-driven decisions. Not only is machine learning becoming increasingly important in computer science research but it also plays an ever greater role in our everyday life. [18]

In Classification Stage, Many machine learning algorithms were used to classify documents based on the feature selected from the classification stage. Machine Learning algorithms are organized into taxonomy, based on the desired outcome of algorithm. Common algorithm type includes: [18]

- **Supervised learning** - where the algorithm generates a function that maps inputs to desire outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn (to approximate the behavior of) a function which maps a vector into one of several classes by looking at several input-output examples of the function.
- **Unsupervised learning** - which models a set of inputs, labeled examples are not available.
- **Semi-supervised learning -** which combines both labeled and unlabeled examples to generate an appropriate function or classifier.
- **Reinforcement learning** - where the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- **Transduction** - similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and new inputs.
- **Learning to learn** where the algorithm learns its own inductive bias based on previous experience.

The performance and computational analysis of machine learning algorithms is a branch of statistics known as computational learning theory. Machine learning is about designing algorithms that allow a computer to learn. Learning is not necessarily involves consciousness but learning is a matter of finding statistical regularities or other patterns in the data. Thus, many machine learning algorithms will barely resemble how human might approach a learning task. However, learning

algorithms can give insight into the relative difficulty of learning in different environments. [18]

2.3.3 Supervised Learning

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete) or a regression function (if the output is continuous). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

The supervised methods can be implemented in variety of domains such as marketing, finance and manufacturing. In order to solve a problem of supervised learning, one has to perform the following steps:

- Determine the type of training examples. Before doing anything else, the engineer should decide what kind of data is to be used as an example. For instance, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.
- Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output
- Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees.

- Complete the design. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.
- Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separated from the training set.

A wide range of supervised learning algorithms is available, each with its strengths and weaknesses. There is no single learning algorithm that works the best on all supervised learning problems. Example of well-known supervised learning algorithm is Decision Tree, Native Bayes, Support Vector Machine and K-Nearest Neighbor and these algorithms are described briefly below.[2]

a) Decision Tree

Decision tree is a flowchart –like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. The construction of the decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. It can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification step of decision tree are simple and fast. In general, decision tree classifier has good accuracy. There are many applications which apply decision tree classifier, such as medicine, manufacturing and production, financial analysis.

b) Native Bayesian Classifier

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is

independent of each other. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. It is easy and fast to predict class of test data set. It also performs well in multi class prediction. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and need less training data. It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption). Naive Bayesian Classifier has been applied in many real world applications.

- **Real Time Prediction:** Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi Class Prediction:** This algorithm is also well known for multi class prediction feature. Here the users can predict the probability of multiple classes of target variable.
- **Text Classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
- **Recommendation System:** Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not

c) Support Vector Machine

In a nutshell, a support vector machine (or SVM) is an algorithm that works as follows. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (that is, a "decision boundary "separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors). Although the training time of even the fastest SVMs

can be extremely slow, they are highly accurate, owing to their ability to model complex nonlinear decision boundaries. They are much less prone to overfitting than other methods. The support vector found also provides a compact description of the learned model. SVMs can be used for prediction as well as classification. They have been applied to number of areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series perdition tests.

d) K-Nearest Neighbor

The K-Nearest Neighbor methods are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all of the training tuples are stored in an n-dimensional space. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuples. These k training tuples are the kth nearest neighbors of the unknown tuple.

Nearest –neighbor classifiers use distance-based comparisons that intrinsically assign equal weight to each attribute. They therefore can suffer from poor accuracy when given noisy or irrelevant attributes. The method, however, has been modified to incorporate attribute weighting and the pruning of noisy data tuples. The choice of a distance metric can be critical.

2.3.4 Unsupervised Learning

Unsupervised learning refers mostly to technique that group instances without a prespecified, dependent attribute. In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning and reinforcement learning .Unsupervised learning is closely related to the problem of density estimation in statistics. However unsupervised learning also encompasses many other techniques that seek to summarize and explain key feature of that data. Many method employed in unsupervised learning is based on

data mining methods used to preprocess data. Approaches to unsupervised learning include:

- **Clustering** – the process of grouping a set of physical or abstract object into classes of similar objects is called clustering. A clustering is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Blind signal separation reduces dimensionality using feature extraction techniques. There are two main clustering techniques namely, hierarchical and partitioned (K-Means) clustering techniques [12]

The unsupervised learning algorithm includes:

(a) Genetic Algorithm

Genetic algorithms are inspired by Darwin's theory about evolution. Solution to a problem solved by genetic algorithms is evolved. It is more suited for finding solutions to complex search problems. They're often used in fields such as engineering to create incredibly high quality products thanks to their ability to search a through a huge combination of parameters to find the best match. For example, they can search through different combinations of materials and designs to find the perfect combination of both which could result in a stronger, lighter and overall, better final product. They can also be used to design computer algorithms, to schedule tasks, and to solve other optimization problems. Genetic algorithms are based on the process of evolution by natural selection which has been observed in nature. They essentially replicate the way in which life uses evolution to find solutions to real world problems. Surprisingly although genetic algorithms can be used to find solutions to incredibly complicated problems, they are themselves pretty simple to use and understand.

The basic process for a genetic algorithm is:

1) **Initialization** - Create an initial population. This population is usually randomly generated and can be any desired size, from only a few individuals to thousands.
2) **Evaluation** - Each member of the population is then evaluated and calculates a 'fitness' for that individual. The fitness value is calculated by how well it fits with the desired requirements. These requirements could be simple, faster algorithms are better, or more complex, stronger materials are better but they shouldn't be too heavy.

3) **Selection** – Population's overall fitness need to be constantly improving by using selection which helps by discarding the bad designs and only keeping the best individuals in the population. There are a few different selection methods but the basic idea is the same, make it more likely that fitter individuals will be selected for our next generation.

4) **Crossover** - During crossover, new individuals were created by combining aspects of the selected individuals. The hope is that by combining certain traits from two or more individuals will create an even 'fitter' offspring which will inherit the best traits from each of its parents.

5) **Mutation** – In mutation, new generation was created which is different from its parent. Mutation typically works by making very small changes at random to an individual's genome.

These processes of Genetic Algorithm repeat until the termination condition met.

(b) Neural Network

An Artificial Neural Network (ANN) is a computational model that is inspired by the way biological neural networks in the human brain process information. Artificial Neural Networks have generated a lot of excitement in Machine Learning research and industry such as speech recognition, computer vision and text processing. It is constructed from a large number of elements with an input in order of magnitudes larger than in computational elements of traditional architectures. [13] These elements, namely artificial neuron are interconnected into group using a mathematical model for information processing based on a connectionist approach to computation. The neural network makes their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors.

Basically, there are three different layers in a neural network:

- Input Layer - All the inputs are fed in the model through this layer.
- Hidden Layers - There can be more than one hidden layers which are used for processing the inputs received from the input layers
- Output Layers - The data after processing is made available at the output layer.

In document classification task, different types of neural network approaches have been implemented. Some of the researches use the single-layer perceptron, which contains only an input layer and an output via a series of weights. In this way, it can be considered as the simplest kind of feed-forward network. The multilayer perception which is more sophisticated, which consists of an input layer, one or more hidden layers, and an output layer in its infrastructure, also widely implemented for classification tasks.

The main benefit of applying the artificial neural network in classification tasks is the ability in handling documents with high-dimensional features, and documents with noisy and contradictory data. The drawback of artificial neural network is its high computing cost which consumes high CPU and physical memory usage. Another limitation is that they were extremely difficult to understand for average users. This may negatively influence the acceptance of these methods.

CHAPTER 3
TEXT CLASSIFIER SYSTEM FOR MYANMAR NEWS ARTICLES

This chapter describes the activities that compose text classification system for Myanmar news articles. It presents details about the nature of Myanmar language, preprocessing techniques for Myanmar Language and applied theories in this system.

3.1 Text Mining and Text Classifier

Nowadays, unstoppable growth of the data from internet or data from business organization demands the comfortable ways to organize and access these data in more easy way. Text mining technique is the most suited ways to solve these problems which can handle and organize 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 text includes text classification, text clustering, entity extraction, and sentiment analysis and text summarization.

Text mining refers to discovery of unknown knowledge or information that can be found in textual data resources. It can be considered as the subtask of data mining field. However, text mining extract patterns from natural language text while data mining does its process by extracting data form database which were well structured. Text mining is the crossroad between the information retrieval and machine learning by applying both methodologies.

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 form text
(c) Extraction of useful and correct information from detected usage patterns

Due to this fact, text classification can be seen as an 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. In this case, training data set is needed.

2) **Unsupervised document classification** (also known as document clustering), where the classification must be done entirely without reference to external information. This type of learning approach the problem without prior knowledge about the problem.

3) **Semi-supervised document classification**, where parts of the documents are labeled by the external mechanism. These learning approaches used the training dataset and approach the problem with little or no idea about what the outcome would be.

Text classification has been applied in many application like text filtering, document organization, classification of news stories, searching for interesting information on the web, spam e-mail filtering etc. Many of these systems based on the specific language and most of these languages were English, European and other Asian Languages and only few works has been done for Myanmar languages. Because of this, classifying Myanmar documents were still facing difficulties as there are morphological richness and security of resources of the languages like automatic tools for tokenization, feature selection and stemming etc.

3.2 Preprocessing In Myanmar Language

Myanmar Language is the official Language of Republic of Union of Myanmar. It can be considered as the Sino-Tibetan family of language of which is a part of Tibetan-Myanmar (Tibeto-Burman) subfamily. Text in the Myanmar language uses the Myanmar script, which derives from a Brahmirelated script borrowed from south India in about the eight centuries for Mon language. The first inscription in Burmese dates form the following years and is written in alphabet almost identical with Mon inscriptions. It is written from left to right and no spaces between words, although informal writing often contains spaces after each clause. Myanmar language

users also normally use space as they see fit or some even write with no space at all. There is no fixed rule for word segmentation.

As there are no strict rules about spaces in sentences, many people write spaces between letters as they feel fit or some didn't put spaces between these words. Myanmar language is a free-word-order and verb final language, which usually follows the subject-object-verb (SOV) order and also sentences contain many preposition adjunctions and appear in several different places.

Myanmar script is composed of 33 consonants, 11 basic vowels, 11 combination symbols and extension vowels, vowel symbols, devowelizing consonants, diacritic marks, specified symbols and punctuation marks [9]. Myanmar has a mainly 9 part of speech: noun, pronoun, verb, adjective, adverb, particle, conjunction, post-positional marker and interjection. [9]

During text mining process, preprocessing of documents from different resources is the necessary step to perform before applying to any text mining technique. Preprocessing text is called tokenization or text normalization. Preprocessing includes word segmentation, stemming, part of speech tagging (POS) and Removing Stop words. Preprocessing stage has been applied based on the nature and necessities of applications. Preprocessing in Myanmar language is still facing difficulties due to language nature and lack of resources. Many researchers have been trying to preprocess raw text using both traditional and statistical methods.

During the words segmentation stage, the words are segmented into valid words. In the task of word segmentation, word tokenization plays an important role in most Natural Language Processing applications. However, in the applied system, word segmentation is done manually .The system only takes already segmented words as an input in the preprocessing process. The word segmentation process words are segmented if the users encounter noun suffix like - များ တို့ တွေ မှု ခြင်း ,also in verb stemming step , the word with suffix like - သည် မည် လိမ့်မည် ကြမည် ပါသည် ကြပါသည် သွားပါသည် ကြောင့် are segmented. Moreover, adj and adv stemming step, words with suffix like သော သည့် မည့် and စွာ can be segmented. These suffixes are considered as stop words and collected for the later used in the classification process.

Stop words removal is removing words that do not contribute the meaning to text mining application from sentences. It should include in all text mining applications to reduce time and memory complexity and to increase performance of the applications.

Stop words are division of natural language. The motive that stop words should be removed from a text is that they make the text look heavier and less important for analysis. Removing stop words reduce the dimensionality of term space. The most common words in text document are articles, prepositions, and pro-nouns. These words are treated as stop words. There is no exact stopword list for Myanmar Language. Stop words are removed from documents because those words are not measured as keywords in text mining applications. Stopword list should be defined by linguists and natural language processing researchers. Stop words for the proposed system are defined and also collected manually while word segmentation stage. Punctuations, special characters and sentence boundary marker are included in this stopword list.

For example, a sentence from the documents can be seen as follow.

“ပြည်နယ်တိုင်းဒေသကြီး ဝန်ကြီးချုပ်များဥက္ကဋ္ဌ အဖြစ် ပါဝင်သည့် ပြည်နယ်နှင့်တိုင်းဒေသကြီး ရင်းနှီးမြှုပ်နှံမှု ကော်မတီ များ ကို မြန်မာနိုင်ငံ ရင်းနှီးမြှုပ်နှံမှု ကော်မရှင် က ဇူလိုင်လ အတွင်း ဖွဲ့စည်း ပေး ခဲ့ ပြီး ဖြစ်ကြောင်း သိရသည်။”

Figure 3.1 A sample sentence from document

A sentence from figure 3.1 is segmented and Figure 3.2 show the segmentation of the sentence.

“ပြည်နယ်-တိုင်းဒေသကြီး-ဝန်ကြီးချုပ်-များ-ဥက္ကဋ္ဌ-အဖြစ်-ပါဝင်-သည့် -ပြည်နယ် -နှင် -တိုင်းဒေသကြီး- ရင်းနှီးမြှုပ်နှံမှုကော်မတီ- များ -ကို- မြန်မာနိုင်ငံ -ရင်းနှီးမြှုပ်နှံမှု ကော်မရှင် -က- ဇူလိုင်လ -အတွင်း -ဖွဲ့စည်း -ပေး -ခဲ့ -ပြီး- ဖြစ်ကြောင်း-သိရသည်”

Figure 3.2 Segmented sample sentence

The stop words are removed according to the collected stop words shown in figure 3.2.

“ပြည်နယ်-တိုင်းဒေသကြီး-ဝန်ကြီးချုပ်-ဥက္ကဋ္ဌ-အဖြစ်-ပါဝင်-ပြည်နယ်-တိုင်းဒေသကြီး-ရင်းနှီးမြှုပ်နှံမှုကော်မတီ-မြန်မာနိုင်ငံ-ရင်းနှီးမြှုပ်နှံမှုကော်မရှင်- ဇူလိုင်လ-အတွင်း-ဖွဲ့စည်း-ပေး”

Figure 3.3 Stopword removal of the sample sentence

3.3 Feature Selection for Text Classification

In the text classification, extracting the right feature for the classification process is important due to the high dimensionality of text features and the existence of irrelevant (noisy) features. In general, text can be represented in two separate ways. The first is as a bag of words, in which a document is represented as a set of words, together with their associated frequency in the documents. Such a representation is essentially independence of the sequence of words in the collection. The second method is to represent text directly as strings, in which each document is a sequence of words. Most text classification methods use the bag-of-words representation because of its simplicity for classification purpose. In this section, some of the methods which are used for feature selection in text classification will be discussed.

The most common feature selection which is used in both supervised and unsupervised applications is that of stop-word removal and stemming. In Stop word removal process, the common words in the documents which are not specific or discriminatory to different classes are determined. In stemming, different form of the same words are consolidated into single words. For example, singular, plural and different tenses are consolidated into a single word. These methods are not specific to the case of the classification problem, and are often used in a variety of unsupervised application such as clustering and indexing. In the case of classification problem, it makes sense to supervise the feature selection process ensure that those features which are highly skewed towards the presence of particular class label are picked for the learning process. [9]

TF-IDF (Term Frequency –Inverse Document Frequency) is used as feature selection algorithm to reduce unnecessary words and to improve the performance of the system. Theoretical background of TF_IDF is already described in Chapter 2.The proposed system took the words that occurred twice in the sentence as

a feature words and calculate the predefined value for them. Typically, the TF_IDF weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. The number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. [25]

The applied system uses labeled documents whose category was already known to use for the training process. Then, calculate the weight of each word. First of all, to calculate the Term Frequency (TF), the system counts the number of the words and total number of words in the same documents. In the case of the term frequency TF, the simplest choice is to use the raw count of a term in a document. Then, it calculates how important the term in its documents is. Secondly, it measures how the importance of the term in the whole corpus. Finally, the system calculates the weight of the term by combining above two measures and store for the later use.

3.4 Genetic Algorithm

The process of natural selection starts with the selection of fittest individuals from a population. They produce offspring which inherit the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance at surviving. This process keeps on iterating and at the end, a generation with the fittest individuals will be found.

This notion can be applied for a search problem. In general, a set of solutions for a problem has been considered and selected the set of best ones out of them.

Five phases are considered in a genetic algorithm.

1) Initial population
2) Fitness function
3) Selection
4) Crossover
5) Mutation

1) Initial Population

The process begins with a set of individuals which is called a Population. Each individual is a solution to the problem that needs to solve. An individual is characterized by a set of parameters (variables) known as Genes. Genes are joined into a string to form a Chromosome (solution).In a genetic algorithm, the set of genes of an individual is represented using a string, in terms of an alphabet. Usually, binary values are used (string of 1s and 0s).It can say that genes are encoded in a chromosome.

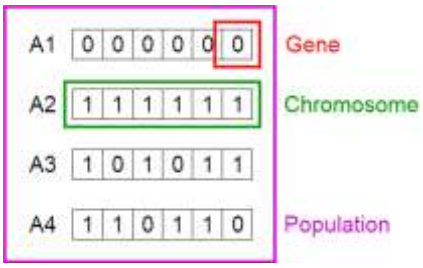

Figure 3.4 Population, Chromosomes and Genes

In the proposed system, value encoding method is used to represent each chromosome. In value encoding, every chromosome is a string of some values. Values can be anything connected to problem; form numbers, real numbers or chars to some complicated objects. Each gene represents predefined categories and its predefined value which is stored during feature selection process. The sample chromosome for the propose system is ခန့်= {e=0.0, s=0.0, b=0.10155024656115223, p=0.12222377773761485}.

2) Fitness Function

The fitness function determines how fit an individual is (the ability of an individual to compete with other individuals). It gives a fitness score to each individual. The probability that an individual will be selected for reproduction is based on its fitness score. As the fitness function is depending on the problem itself .The applied system uses the standard deviation to calculate the fitness value for each feature .WTSD (weight Term Standard Derivation) fitness function.

$$\textbf{WTSD } (d_i) = \sum \frac{wd_i c_j (x_{j,k} - \bar{x}.wd_i c_j)^2}{(n-1).\bar{w}.maxfrequent(d_i)} \qquad \text{Eq [3.1]}$$

In Equitation [3.1] ,d_i is document that is 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) Selection

The idea of selection phase is to select the fittest individuals and let them pass their genes to the next generation. Two pairs of individuals (parents) are selected based on their fitness scores. Individuals with high fitness have more chance to be selected for reproduction.

Among the many selection method, the propose system used the tournament selection method to select two fittest parents. Tournament selection is a method of selecting an individual from a population of individuals in a genetic algorithm. Tournament selection involves running several "tournaments" among a few individuals (or "chromosomes") chosen at random from the population. The winner of each tournament (the one with the best fitness) is selected for crossover. If the tournament size is larger, weak individuals have a smaller chance to be selected because if a weak individual is selected to be in a tournament, there is a higher probability that a stronger individual is also in that tournament.

Crossover and mutation are two basic operators of GA. Performance of GA very depend on them. Type and implementation of operators depends on encoding and also on a problem. This proposed genetic algorithm will apply binary encoding method for both crossover and mutation.

4) Crossover

Crossover is the most significant phase in a genetic algorithm. For each pair of parents to be mated, a crossover point is chosen at random from within the genes. The propose system will use the single point crossover method. Single point crossover can be defined as one crossover point is selected, binary string from beginning of chromosome to the crossover point is copied from one parent, and the rest is copied from the second parent.

5) Mutation

In certain new offspring formed, some of their genes can be subjected to a mutation with a low random probability. This implies that some of the bits in the bit string can be flipped. The system will be used Bit inversion method to invert the selected bits. Mutation occurs to maintain diversity within the population and prevent premature convergence.

6) Termination

The algorithm terminates if the population has converged (does not produce offspring which are significantly different from the previous generation). Then it is said that the genetic algorithm has provided a set of solutions to the problem.

3.5 Algorithm of Text Classification System for Myanmar News Articles

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 documents to classify these into their related category.

Figure 3.5 - Proposed Algorithm for Myanmar Text Classifier using Genetic algorithm

3.6 Example Classification Using Genetic Algorithm

A Step by Step detail calculation of algorithm of text classification system is described in this section by using an example of small dataset.

3.6.1 Dataset

There is a total of 11 documents in an example dataset will be used as training documents and these dataset are collected from online news documents written in Myanmar Language text. In preprocessing stage, word segmentation is done in manually. The propose system only takes already segmented text as an input. Among these 11 documents, p1.txt will be used to explain the preprocessing stage. The labels of these 11 documents are already known and after the calculation of the weight, the words are store as weight or predefined value for their related category. The proposed system is going to classify documents into four categories such as Politic, Business, Entertainment and sport. In the following example, the documents 'label is politic and the calculated weight will be stored as feature for politic.

P1.txt

မြန်မာ အမျိုးသမီးတွေ အိမ်တွင်း အကြမ်းဖက် ခံရမှု ပိုများလာ

မြန်မာနိုင်ငံ နယ်စပ် ဒေသတွေမှာ ဖြစ်ပွားနေတဲ့ စစ်မက် ဆိုးကျိုး တွေ၊ နိုင်ငံ တလွှားက အလုပ်လက်မဲ့ ပြဿနာ၊ မူးယစ်ဆေး ကျယ်ကျယ်ပြန့်ပြန့် ရောင်းဝယ်မှုနဲ့ တခြား လူမှုဒုက္ခ တွေကြောင့် အမျိုးသမီးတွေရဲ့ ဘဝလုံခြုံမှု ကင်းမဲ့လာနေပြီး အိမ်တွင်း အကြမ်းဖက်မှုဒဏ် ပိုမို ဆိုးဆိုးရွားရွား ခံလာရတယ်လို့ မြန်မာနိုင်ငံ အမျိုးသမီးများ အဖွဲ့ချုပ်က ထုတ်ပြန်တဲ့ ကြေညာချက်မှာ ဖော်ပြထားတယ် လို့ဒုဥက္ကဋ္ဌ ဒေါ်စောစံ ငြိမ်းသူ က ပြောပါတယ်။

Figure 3.6 Example Document

3.6.2 Training Dataset

The sample document p1.txt from figure 3.6 are segmented manually as to prepare for the input into the documents and segmented according to the segmentation rules. The P1.txt is segmented into following figure 3.7.

မြန်မာ -အမျိုးသမီး- တွေ - အိမ်တွင်း -အကြမ်းဖက် - ခံ- ရ - မှု - ပို -များ - လာ - မြန်မာနိုင်ငံ - နယ်စပ် - ဒေသ - တွေ - မှာ - ဖြစ်ပွား - နေ - တဲ့ - စစ်မက် - ဆိုးကျိုး - တွေ - နိုင်ငံ - တလွား - က -အလုပ်လက်မဲ့ - ပြဿနာ - မူးယစ်ဆေး - ကျယ်ကျယ်ပြန့်ပြန့် - ရောင်းဝယ် - မှု - နဲ့ - တခြား - လူမှု - ဒုက္ခ - တွေ - ကြောင့် - အမျိုးသမီး - တွေ - ရဲ့- ဘဝ - လုံခြုံ - မှု - ကင်းမဲ့ - လာ - နေ - ပြီး - အိမ်တွင်း - အကြမ်းဖက် - မှု - ဒဏ် - ပိုမို - ဆိုးဆိုးရွားရွား - ခံ - လာ - ရ - တယ် - လို့ - မြန်မာနိုင်ငံ - အမျိုးသမီး - များ - အဖွဲ့ချုပ် - က - ထုတ်ပြန် - တဲ့ - ကြေညာ - ချက် - မှာ - ဖော်ပြ - ထား - တယ် - လို့ - ဒုဥက္ကဋ္ဌ - ဒေါ်စောစံ ငြိမ်းသူ - က - ပြော - ပါတယ်

Figure 3.7 Segmentation of Sample Document

Stop Words are removed from document and the users can store in the stop words Collection file and can also use in classification stage.

မြန်မာ -အမျိုးသမီး- အိမ်တွင်း -အကြမ်းဖက် - ခံ - မြန်မာနိုင်ငံ - နယ်စပ် - ဒေသ - ဖြစ်ပွား - စစ်မက် - ဆိုးကျိုး - နိုင်ငံ တလွား -အလုပ်လက်မဲ့ - ပြဿနာ - မူးယစ်ဆေး - ကျယ်ကျယ်ပြန့်ပြန့် - ရောင်းဝယ် - လူမှု - ဒုက္ခ - အမျိုးသမီး - ဘဝ - လုံခြုံ - ကင်းမဲ့ - အိမ်တွင်း - အကြမ်းဖက် - ဒဏ် - ဆိုးဆိုးရွားရွား - မြန်မာနိုင်ငံ - အမျိုးသမီး - အဖွဲ့ချုပ် - ထုတ်ပြန် - ကြေညာ - ချက်- ဖော်ပြ - ဒုဥက္ကဋ္ဌ - ဒေါ်စောစံ ငြိမ်းသူ - ပြော

Figure 3.8 Removing Stop words from Sample Document

Within these words, the users only collect words whose frequency is greater than 2 in predefined document as feature to train the algorithm.

Term
မြန်မာနိုင်ငံ
အမျိုးသမီး
အိမ်တွင်း
အကြမ်းဖက်

Figure 3.9 Collected Terms from Sample Document

Rest of the document will be used in the preprocessing stage to segment text, remove stop words and to use collection of these stop words and calculate TF_IDF (term frequency and inverse document frequency) feature selection method. The term frequency (TF) is the number of times and a term (word) occurs in a document.

$$\mathbf{TF} = \frac{\text{Number of times terms t appear in a document}}{\text{total number of terms in the documents}} \qquad \text{Eq[3.2]}$$

Inverse Document frequency (IDF) measures how important a term is for the corpus.

$$\mathbf{IDF} = \log\left(\frac{\text{total number of document}}{\text{total number of documenttotal number of document}}\right) \qquad \text{Eq [3.3]}$$

The concepts of term frequency and inverse document frequency are combined to produce a composite weight for each term in each document.

$$\boldsymbol{TF - IDF} = TF * IDF \qquad \text{Eq [3.4]}$$

The following calculation shows how each feature predefined value is calculated:

Total no of document (all document which were used in training) - 11

Total no of term in a document - 37

Number of times terms t appear in a document - 2

Number of documents with term t in it-3(among 11 documents)

Calculation of TF-IDF for words- မြန်မာနိုင်ငံ

TF-IDF (မြန်မာနိုင်ငံ)

TF=$\frac{2}{37}$=0.0540540541

IDF =log($\frac{11}{3}$)=0.5642

TF-IDF =TF*IDF =0.0304972

Term	TF	IDF	TF-IDF
မြန်မာနိုင်ငံ	0.0540540541	0.5642	0.0304972
အမျိုးသမီး	0.081081	0.7403	0.06002
အိမ်တွင်း	0.0540540541	0.5642	0.0304972
အကြမ်းဖက်	0.0540540541	0.5642	0.0304

Table 3.1 Calculation of TF-IDF

After calculating the TF-IDF method for predefined process, terms are restored by genetic algorithm in classification stage .Collected feature words are as follow.

term	P1	P2	P3	P4	P5	P6	P7	P8
မြန်မာနိုင်ငံ	0.0304972				0.0497	0.0110		
အမျိုးသမီး	0.06002	0.0860						
အိမ်တွင်း	0.0304972							
အကြမ်းဖက်	0.0304972							
ငြိမ်းချမ်း		0.0725						
ဖြစ်စဉ်		0.0484						
တိုက်တွန်း		0.0484						
နေ့		0.0484						
အတိုင်ပင်ခံပုဂ္ဂိုလ်		0.0484						
ပါဝင်		0.0368						
သမ္မတ			0.0155					
ရုံး			0.0218					
ခွင့်			0.0218					
လောက်ကိုင်			0.0218					

မြို့			0.0218				0.0321	
တိုက်ပွဲ			0.0218					
လက်နက်			0.0218					
ကိုင်			0.0218					
ဦးဇော်ဌေး			0.0438					
ကိုးကန့်			0.0438					
ပြော			0.0328					
အဖွဲ့			0.0233					
တရား				0.0457				
ကဏ္ဍ				0.0228				
တည့်မတ်				0.0228				
အတည်ပြု				0.0228				
ပြည်သူ				0.0228				
ဆွေးနွေး				0.0228				
လွှတ်တော်				0.0800				
မဏ္ဍိုင်				0.0342				
အဆို				0.0342				
နိုင်ငံသား					0.0306			
အသိအမှတ်					0.0204			
တရားမဝင်					0.0204			
အစိုးရ					0.0204			
လက်မှတ်					0.0588			
ထိုင်း					0.0588			
အလုပ်သမား					0.0510			

ကချင်						0.0306		
ရိုဟင်ဂျာ						0.0306		
လူ့အခွင့်အရေး						0.0204		
ခုံရုံး						0.0204		
ကြားနာ						0.0306		
သက်သေ						0.0306		
ရန်ကုန်							0.0213	
လေထု							0.0753	
ညစ်ညမ်း							0.0753	
လျှော့ချ							0.0753	
နိုင်							0.0602	
ဦးကိုနီ								0.0328
တရားစွဲ								0.0328
ဥပဒေ								0.0328
အမှု								0.0328
အဖွဲ့								0.0233
ပစ်သတ်								0.0218
ကြည်လင်း								0.0218
ပုဒ်မ								0.0218
အထောက်အကူ								0.0218
ဦးကျော်ဟိုး								0.0218

Table 3.2 Collected Feature Words

After Calculating the weight of each term for each feature , the feature with highest weight are selected as a weight of the feature or predefined value for each feature. These values are stored for the classification stage.

3.6.3 Classification of Documents

During the classification stage, unknown documents with unknown label are used to classify into the already predefined category.

"ပြည်ထောင်စု ငြိမ်းချမ်းရေးညီလာခံ-၂၁ ရာစုပင်လုံ ဒုတိယအစည်းအဝေးအား မေ ၁၉ ရက်တွင် စတင်ရန် အစိုးရနှင့် NCA လက်မှတ်ရေးထိုးထားသည့် တိုင်းရင်းသား လက်နက်ကိုင်အဖွဲ့အစည်းများညှိနှိုင်းလျာထား၂၀၁၆ခုနှစ်ပြည်ထောင်စုငြိမ်းချမ်းရေးညီလာခံ-၂၁ ရာစုပင်လုံ ပထမအစည်းအဝေးအား နေပြည်တော်တွင် ပြုလုပ်ခဲ့သည်ကို တွေ့ရစဉ် ပြည်ထောင်စု ငြိမ်းချမ်းရေးညီလာခံ -၂၁ ရာစုပင်လုံ ဒုတိယအစည်းအဝေးအားမေ၁၉ ရက်တွင် စတင်ပြုလုပ်ရန် အစိုးရနှင့် တစ်နိုင်ငံလုံး ပစ်ခတ်တိုက်ခိုက်မှု ရပ်စဲရေး သဘောတူစာချုပ် (NCA) လက်မှတ်ရေးထိုးထားသည့် တိုင်းရင်းသား လက်နက်ကိုင်အဖွဲ့အစည်းများ ညှိနှိုင်းလျာထားကြောင်း ပအိုဝ်းတိုင်းရင်းသား လက်နက်ကိုင်အဖွဲ့ (PNLO) နာယက ဗိုလ်မှူးကြီး ခွန်ဥက္ကာက ဧပြီ ၂၂ ရက် တွင် ပြောကြား သည်။ညီလာခံ ၏ ပထမအစည်းအဝေးအား စက်တင်ဘာ လဆန်းပိုင်း တွင် ကျင်းပပြီးစီးခဲ့ပြီး ဒုတိယအစည်းအဝေးကိုမူ ပုံမှန်အားဖြင့် ဖေဖော်ဝါရီလအတွင်းကျင်းပရမည် ဖြစ်သော်လည်း ကျန်ရှိနေသည့် အမျိုးသားအဆင့် နိုင်ငံရေးဆွေးနွေးပွဲများအား စောင့်ဆိုင်းခြင်း၊ အမျိုးသားအဆင့် နိုင်ငံရေး ဆွေးနွေးပွဲ များ ၏ နောက်ဆက်တွဲ လုပ်ငန်းစဉ်များ လုပ်ဆောင်ရန် ကျန်ရှိနေခြင်း၊ NCA လက်မှတ်ရေးထိုးထားခြင်း မရှိသေးသည့် တိုင်းရင်းသား လက်နက်ကိုင်အဖွဲ့များ ပါဝင်နိုင်ရန်အတွက် စောင့်ဆိုင်းခြင်းတို့ကြောင့် မေလအတွင်းသို့ ရွှေ့ဆိုင်းခဲ့သည်။"

Figure 3.10 Input Document

This unknown document is segmented into following and considered as an input into the system.

"ပြည်ထောင်စု- ငြိမ်းချမ်း -ရေး-ညီလာခံ-၂၁ -ရာစု-ပင်လုံ- ဒုတိယ-အစည်းအဝေး-အား-မေ- ၁၉ -ရက်-တွင် -စတင်-ရန်- အစိုးရ-နှင့်- NCA -လက်မှတ်-ရေးထိုး-ထား-သည့်-

တိုင်းရင်းသား -လက်နက်-ကိုင် -အဖွဲ့အစည်း-များ- ညှိနှိုင်း-လျာထား-၂၀၁၆- ခုနှစ်-ပြည်ထောင်စု-ငြိမ်းချမ်းရေး-ညီလာခံ-၂၁-ရာစု-ပင်လုံ-ပထမ-အစည်းအဝေး-အား-
နေပြည်တော်-တွင် -ပြုလုပ်-ခဲ့-သည်-ကို- တွေ့ရ-စဉ်-ပြည်ထောင်စု- ငြိမ်းချမ်း-ရေး-ညီလာခံ -၂၁- ရာစု-ပင်လုံ- ဒုတိယ-အစည်းအဝေး-အား- မေ -၁၉ -ရက်-တွင်- စတင်-ပြုလုပ်-ရန် -အစိုးရ-နှင့်-တစ်-နိုင်ငံ-လုံး-ပစ်ခတ်-တိုက်ခိုက်-မှု-ရပ်စဲ-ရေး-သဘောတူ-စာချုပ်-လက်မှတ်-ရေးထိုး-ထား-သည့်-တိုင်းရင်းသား-လက်နက်-ကိုင်- အဖွဲ့အစည်း-များ- ညှိနှိုင်း-လျာ-ထား-ကြောင်း- ပအိုဝ်း-တိုင်းရင်းသား- လက်နက်-ကိုင်-အဖွဲ့- နာယက -ဗိုလ်မှူးကြီး -ခွန်ဥက္ကာ-က -ဧပြီ -၂၂ -ရက်-တွင်- ပြောကြား-သည်-ညီလာခံ-၏-ပထမ-အစည်းအဝေး-အား-စက်တင်ဘာ -လဆန်း-ပိုင်း-တွင်- ကျင်းပ-ပြီးစီး-ခဲ့-ပြီး-ဒုတိယ-အစည်းအဝေး-ကို-မူ-ပုံမှန်-အား-ဖြင့် -ဖေဖော်ဝါရီ-လ-အတွင်း -ကျင်းပ-ရမည်- ဖြစ်-သော်လည်း- ကျန်ရှိ-နေ-သည့် -အမျိုးသား-အဆင့်-နိုင်ငံရေး-ဆွေးနွေး-ပွဲ-များ-အား-စောင့်ဆိုင်း-ခြင်း- အမျိုးသား-အဆင့်-နိုင်ငံ-ရေး-ဆွေးနွေး-ပွဲ-များ-၏ -နောက်ဆက်-တွဲ- လုပ်ငန်း-စဉ်-များ -လုပ်ဆောင်-ရန်-ကျန်ရှိ-နေ-ခြင်း-လက်မှတ်-ရေးထိုး-ထား-ခြင်း-မ-ရှိ-သေး-သည့်-တိုင်းရင်းသား-
လက်နက်-ကိုင်-အဖွဲ့-များ- ပါဝင်-နိုင်-ရန်-အတွက် စောင့်ဆိုင်း-ခြင်း-တို့-ကြောင့်- မေ-လ-အတွင်း-သို့ -ရွှေ့ဆိုင်း-ခဲ့-သည်”

Figure 3.11 Segmentation of Input Document

The system removes the stop words from the previous collected stopword list.

ပြည်ထောင်စု- ငြိမ်းချမ်း -ညီလာခံ-ရာစု-ပင်လုံ- ဒုတိယ-အစည်းအဝေး-စတင်- အစိုးရ-လက်မှတ်-ရေးထိုး- တိုင်းရင်းသား -လက်နက်-ကိုင် -အဖွဲ့အစည်း- ညှိနှိုင်း-လျာထား-ခုနှစ်- ပြည်ထောင်စု- ငြိမ်းချမ်း-ညီလာခံ- ရာစု-ပင်လုံ- ပထမ-အစည်းအဝေး-နေပြည်တော်-ပြုလုပ်- တွေ့ရ- ပြည်ထောင်စု- ငြိမ်းချမ်း -ညီလာခံ -၂၁- ရာစု-ပင်လုံ-ဒုတိယ-အစည်းအဝေး- မေ - စတင်-ပြုလုပ်-အစိုးရ- တစ်-နိုင်ငံ-ပစ်ခတ်-တိုက်ခိုက်- ရပ်စဲ-သဘောတူ-စာချုပ်-လက်မှတ်-ရေးထိုး- တိုင်းရင်းသား -လက်နက်-ကိုင်- အဖွဲ့အစည်း-

ညှိနှိုင်း-လျာ- ပအိုဝ်း-တိုင်းရင်းသား- လက်နက်-ကိုင်-အဖွဲ့- နာယက -ဗိုလ်မှူးကြီး - ခွန်ဥက္ကာ- ပြောကြား-ညီလာခံ-ပထမ-အစည်းအဝေး-စက်တင်ဘာ -လဆန်း- ကျင်းပ-ပြီးစီး-ဒုတိယ-အစည်းအဝေး- ပုံမှန်-ဖေဖော်ဝါရီ-လ- ကျင်းပ- ကျန်ရှိ-အမျိုးသား-အဆင့် - နိုင်ငံ-ဆွေးနွေး-ပွဲ- စောင့်ဆိုင်း- အမျိုးသား-အဆင့်- နိုင်ငံ-ဆွေးနွေး-ပွဲ-နောက်ဆက်-တွဲ-လုပ်ငန်း- လုပ်ဆောင်- ကျန်ရှိ- လက်မှတ်-ရေးထိုး- တိုင်းရင်းသား -လက်နက်-ကိုင်-အဖွဲ့-ပါဝင်- အတွက် စောင့်ဆိုင်း- မေ-လ-ရွေ့ဆိုင်း

Figure 3.12 Removal of Stopword form Input Document

The predefined document words are collected in the earlier stage and these collected words are used as the initial population.

Politic	**Business**	**Entertainment**	**Sport**
ပြည်ထောင်စု(p1) ငြိမ်းချမ်း(p1,p4,p6,p7) ညီလာခံ(p1,p4) ရာစု(p1,p4) ပင်လုံ(p1,p4) ဒုတိယ(p1) အစည်းအဝေး(p1,p4) အစိုးရ(p2,p3,p5,p7,p8,p9,p10) လက်မှတ်(p5) တိုင်းရင်းသား(p6,p5) လက်နက်(p5) ကိုင်(p5) အဖွဲ့အစည်း(p8) ညှိနှိုင်း(p4)	လုပ်ငန်း(B1,B2,B6) ပြောကြား(B4,B9) ကျင်းပ(B5) ခုနှစ်(B9)	အဖွဲ့(E8,E3) ပွဲ(E8) ပြုလုပ်(E3,E6) နိုင်ငံ(E9)	----not found---

ပြုလုပ်(p1,p4,p9) နိုင်ငံ(p1,p2,p4,p7,p9) စာချုပ်(p5) အဖွဲ့(p10,p8,p7,p4,p5) ပြောကြား(p9,p7,p6,p2,p1) ကျင်းပ(p7,p4) အမျိုးသား(p7,p4) ဆွေးနွေး(p1,p7,p9,p4) ပွဲ(p1,p4,[7) လုပ်ငန်း(p10,p8,p4)			

Table 3.3 Collected Terms from the Input Document

Individual [politics, business, entertainment, sport]

Term	**Chromosome**
ပြည်ထောင်စု	[0.0476,0,0,0]
ငြိမ်းချမ်း	[0.0568,0,0,0]
ညီလာခံ	[0.0436,0,0,0]
ရာစု	[0.0443,0,0,0]
ပင်လုံ	[0.0443,0,0,0]
ဒုတိယ	[0.0476,0,0,0]
အစည်းအဝေး	[0.01310,0,0,0]
အစိုးရ	[0.0072,0,0,0]
လက်မှတ်	[0.0270,0,0,0]
တိုင်းရင်းသား	[0.1854,0,0,0]
လက်နက်	[0.0540,0,0,0]

ကိုင်	[0.0540,0,0,0]
အဖွဲ့အစည်း	[0.0303,0,0,0]
ညှိနှိုင်း	[0.0833,0,0,0]
ပြုလုပ်	[0.1041,0,0.0279,0]
နိုင်ငံ	[0.0158,0.1307,0.0428,0]
စာချုပ်	[0.0270,0,0,0]
အဖွဲ့	[0.0243,0,0.0249,0]
ပြောကြား	[0.0209,0.0256,0,0]
ကျင်းပ	[0.0517,0.0256,0,0]
အမျိုးသား	[0.0436,0,0,0]
ဆွေးနွေး	[0.0746,0,0,0]
ပွဲ	[0.0645,0,0.0757,0]
လုပ်ငန်း	[0.04355,0.0677,0,0]
ခုနှစ်	[0,0.0338,0,0]

Table 3.4 Terms and Chromosome

After collecting feature words, the users initialize the population. Among these chromosomes (individual), weight from each category is compared to find the greatest which is considered as weight of each individual.

No	**Term**	**Chromosome**	**Frequency**
1	ပြည်ထောင်စု	[0.0476,0,0,0]	3
2	ငြိမ်းချမ်း	[0.0568,0,0,0]	3
3	ညီလာခံ	[0.0436,0,0,0]	4
4	ရာစု	[0.0443,0,0,0]	3
5	ပင်လုံ	[0.0443,0,0,0]	3

6	ဒုတိယ	[0.0476,0,0,0]	3
7	အစည်းအဝေး	[0.01310,0,0,0]	5
8	အစိုးရ	[0.0072,0,0,0]	2
9	လက်မှတ်	[0.0270,0,0,0]	3
10	တိုင်းရင်းသား	[0.1854,0,0,0]	3
11	လက်နက်	[0.0540,0,0,0]	3
12	ကိုင်	[0.0540,0,0,0]	3
13	အဖွဲ့အစည်း	[0.0303,0,0,0]	1
14	ညှိနှိုင်း	[0.0833,0,0,0]	1
15	ပြုလုပ်	[0.1041,0,0.0279,0]	2
16	နိုင်ငံ	[0.0158,0.1307,0.0428,0]	3
17	စာချုပ်	[0.0270,0,0,0]	1
18	အဖွဲ့	[0.0243,0,0.0249,0]	2
19	ပြောကြား	[0.0209,0.0256,0,0]	1
20	ကျင်းပ	[0.0517,0.0256,0,0]	2
21	အမျိုးသား	[0.0436,0,0,0]	2
22	ဆွေးနွေး	[0.0746,0,0,0]	2
23	ပွဲ	[0.0645,0,0.0757,0]	2
24	လုပ်ငန်း	[0.04355,0.0677,0,0]	1
25	ခုနှစ်	[0,0.0338,0,0]	1

Table 3.5 Tem with each frequency and chromosome

Calculating fitness function

Calculating fitness for each words

$$\text{WTSD } (d_i) = \sqrt{\sum \frac{wd_i c_j (x_{j,k} - \bar{x}.wd_i c_j)^2}{(n-1).\bar{w}.maxfrequent(d_i)}} \quad \text{Eq [3.5]}$$

d_i = document that were classified

w_{dicj}= weight of the term

$x_{j,k}$=number of frequent that term j accursed

n= total number of terms in a document

$\bar{x}$ =mean of frequent of the term

$\bar{w}$ =average of weight of term in document

No	Term	Chromosome	weight	Frequency	Fitness value
1	ပြည်ထောင်စု	[0.0476,0,0,0]	0.0476	3	0.2432
2	ငြိမ်းချမ်း	[0.0568,0,0,0]	0.0568	3	0.2636
3	ညီလာခံ	[0.0436,0,0,0]	0.0436	4	0.3141
4	ရာစု	[0.0443,0,0,0]	0.0443	3	0.2352
5	ပင်လုံ	[0.0443,0,0,0]	0.0443	3	0.2352
6	ဒုတိယ	[0.0476,0,0,0]	0.0476	3	0.2432
7	အစည်းအဝေး	[0.01310,0,0,0]	0.01310	5	0.2195
8	အစိုးရ	[0.0072,0,0,0]	0.0072	2	0.0649
9	လက်မှတ်	[0.0270,0,0,0]	0.0270	3	0.1862
10	တိုင်းရင်းသား	[0.1854,0,0,0]	0.1854	3	0.4259
11	လက်နက်	[0.0540,0,0,0]	0.0540	3	0.2576
12	ကိုင်	[0.0540,0,0,0]	0.0540	3	0.2576

13	အဖွဲ့အစည်း	[0.0303,0,0,0]	0.0303	1	0.0623
14	ညှိနှိုင်း	[0.0833,0,0,0]	0.0833	1	0.0895
15	ပြုလုပ်	[0.1041,0,0.0279,0]	0.1041	2	0.2185
16	နိုင်ငံ	[0.0158,0.1307,0.0428,0]	0.1307	3	0.3756
17	စာချုပ်	[0.0270,0,0,0]	0.0270	1	0.0593
18	အဖွဲ့	[0.0243,0,0.0249,0]	0.0249	2	0.1182
19	ပြောကြား	[0.0209,0.0256,0,0]	0.0256	1	0.0580
20	ကျင်းပ	[0.0517,0.0256,0,0]	0.0517	2	0.1641
21	အမျိုးသား	[0.0436,0,0,0]	0.0436	2	0.1529
22	ဆွေးနွေး	[0.0746,0,0,0]	0.0746	2	0.1923
23	ပွဲ	[0.0645,0,0.0757,0]	0.0757	2	0.1934
24	လုပ်ငန်း	[0.04355,0.0677,0,0]	0.0677	1	0.0843
25	ခုနှစ်	[0,0.0338,0,0]	0.0338	1	0.0653

Table 3.6 Fitness Value for Each Term

Creating new empty child population C to store the new offspring with the same number as the initial parent population. While not enough individual in C{while size(C<N) .i.e. N=initial population=25}

Select two parents for crossover operation (using tournament selection method)

```
begin
function tournament _selection (pop , k )
        Best=null;
        for i=1 to k
                ind=[random(1,N)]
                If (best=null) or fitness (ind) > fitness (best)
                        Best = ind
        return best
```

end

Figure 3.13 Algorithm for Tournament Selection

Selection

After using tournament selection method for the selection of two parents, the selections of first parent return 5 and selection of second return 13.

Parent 1: chromosome [5]➔ [0.0303, 0, 0, 0]

Parent 2: chromosome [13]➔ [0.0303, 0, 0, 0]

Crossover the selected parent to form new offspring

Random [1,4]=3

Parent 1: chromosome [5]➔ [0.0303, 0, 0, 0]

Parent 2: chromosome [13]➔ [0.0303, 0, 0, 0]

After crossover

New offspring= [0.0303, 0, 0, 0]

Mutation

Random [1, 4] =2

New offspring= [0.0303, 0.01, 0, 0]

The iteration will continue for all the feature words if iteration is finished. The result of the child population is as follow.

After finishing the iteration, the result of child population is as follow

No	**Term**	**Chromosome**
1	ပြည်ထောင်စု	[0.0303,0.01,0,0]
2	ငြိမ်းချမ်း	[0.0476,0,0,0.01]
3	ညီလာခံ	[0.0933,0,0,0]
4	ရာစု	[0.0517,0,0,0.01]
5	ပင်လုံ	[0.01310,0,0,0.01]
6	ဒုတိယ	[0.0476,0.01,0,0]
7	အစည်းအဝေး	[0.0640,0,0,0]

8	အစိုးရ	[0.0258,0.1307,0.0428,0]
9	လက်မှတ်	[0.0540,0,0.0279,0.01]
10	တိုင်းရင်းသား	[0.0476,0,0.01,0]
11	လက်နက်	[0.0540,0,0.01,0]
12	ကိုင်	[0.1854,0,0,0.01]
13	အဖွဲ့အစည်း	[0.0746,0,0,0.01]
14	ညှိနှိုင်း	[0.1141,0,0,0]
15	ပြုလုပ်	[0.0517,0.0256,0,0.01]
16	နိုင်ငံ	[0. 01310,0.01,0,0]
17	စာချုပ်	[0.0833,0,0,0.01]
18	အဖွဲ့	[0.0933,0,0,0]
19	ပြောကြား	[0.0158,0.1407,0.0428,0]
20	ကျင်းပ	[0.0476,0,0.01,0]
21	အမျိုးသား	[0.0568,0.01,0,0]
22	ဆွေးနွေး	[0.1854,0.01,0,0]
23	ပွဲ	[0.0476,0.01,0,0]
24	လုပ်ငန်း	[0.0443,0,0,0.01]
25	ခုနှစ်	[0.0476,0.01,0,0]

Table 3.7 Child Population

The child population has been evaluated to use as the population for next generation.

No	Term	Chromosome	weight	Frequency	Fitness value
1	ပြည်ထောင်စု	[0.0303,0.01,0,0]	0.0303	3	0.1724
2	ငြိမ်းချမ်း	[0.0476,0,0,0.01]	0.0476	3	0.2131

3	ညီလာခံ	[0.0933,0,0,0]	0.0933	4	0.3906
4	ရာစု	[0.0517,0,0,0.01]	0.0517	3	0.2214
5	ပင်လုံ	[0.01310,0,0,0.01]	0.0131	3	0.1149
6	ဒုတိယ	[0.0476,0.01,0,0]	0.0476	3	0.2131
7	အစည်းအဝေး	[0.0640,0,0,0]	0.0640	5	0.4151
8	အစိုးရ	[0.0258,0.1307,0.0428,0]	0.1307	2	0.2069
9	လက်မှတ်	[0.0540,0,0.0279,0.01]	0.0540	3	0.2258
10	တိုင်းရင်းသား	[0.0476,0,0.01,0]	0.0476	3	0.2131
11	လက်နက်	[0.0540,0,0.01,0]	0.0540	3	0.2258
12	ကိုင်	[0.1854,0,0,0.01]	0.1854	3	0.3733
13	အဖွဲ့အစည်း	[0.0746,0,0,0.01]	0.0746	1	0.0761
14	ညှိနှိုင်း	[0.1141,0,0,0]	0.1141	1	0.0835
15	ပြုလုပ်	[0.0517,0.0256,0,0.01]	0.0517	2	0.1444
16	နိုင်ငံ	[0. 01310,0.01,0,0]	0.0131	3	0.1149
17	စာချုပ်	[0.0833,0,0,0.01]	0.0833	1	0.0784
18	အဖွဲ့	[0.0933,0,0,0]	0.0933	2	0.1839
19	ပြောကြား	[0.0158,0.1407,0.0428,0]	0.1407	1	0.0847
20	ကျင်းပ	[0.0476,0,0.01,0]	0.0476	2	0.1393
21	အမျိုးသား	[0.0568,0.01,0,0]	0.0568	2	0.1504
22	ဆွေးနွေး	[0.1854,0.01,0,0]	0.1854	2	0.2276
23	ပွဲ	[0.0476,0.01,0,0]	0.0476	2	0.1393
24	လုပ်ငန်း	[0.0443,0,0,0.01]	0.0443	1	0.0637
25	ခုနှစ်	[0.0476,0.01,0,0]	0.0476	1	0.0655

Table 3.8 2nd Generation child chromosome

After calculate selection, crossover, mutation for all 25 chromosomes (individual), and the result of 3rd generation child chromosomes are as follow.

No	Term	Chromosome
1	ပြည်ထောင်စု	[0.1854,0.02,0.01,0.01]
2	ငြိမ်းချမ်း	[0. 0640,0.01,0,0]
3	ညီလာခံ	[0. 02310,0.02,0,0.01]
4	ရာစု	[0.0576,0.01,0.01,0]
5	ပင်လုံ	[0.0517,0.01,0.01,0.01]
6	ဒုတိယ	[0. 0540, 0.01,0.02,0]
7	အစည်းအဝေး	[0.0476,0,0.02,0]
8	အစိုးရ	[0.1033,0, 0.01,0.02]
9	လက်မှတ်	[0.0617,0.01,0,0]
10	တိုင်းရင်းသား	[0.0640,0.01,0.01,0]
11	လက်နက်	[0.0640,0,0, 0.01]
12	ကိုင်	[0. 03310,0.01,0,0.01]
13	အဖွဲ့အစည်း	[0.1854,0.01,0.01,0.02]
14	ညှိနှိုင်း	[0.0640,0,0,0]
15	ပြုလုပ်	[0.1954,0,0.0379,0.01]
16	နိုင်ငံ	[0.0540,0,0.01,0.01]
17	စာချုပ်	[0.0517,0.0256,0.01,0.02]
18	အဖွဲ့	[0. 0476,0,0.02,0.01]
19	ပြောကြား	[0. 1854,0.01,0.02,0]
20	ကျင်းပ	[0.0740,0.01,0,0]
21	အမျိုးသား	[0.0540,0.01,0.02,0]
22	ဆွေးနွေး	[0. 1854,0.01,0.02,0]

23	ပွဲ	[0.1854, 0.01,0.02,0.01]
24	လုပ်ငန်း	[0.0640,0, 0.01,0]
25	ခုနှစ်	[0. 0540,0,0.0479,0.01]

Table 3.9 3rd Generation child chromosome

After calculation, each individual's political chromosome is greater than other chromosome. So, the system can finally conclude that the predicted category of this input document is Politic.

CHAPTER 4

DESIGN AND IMPLEMENTATION OF THE SYSTEM

This chapter describes the design, implementation, experimental result of the system. The system is implemented with java programming language and uses Myanmar 3 font for the textual data. The system is experimented on the dataset available from Myanmar news websites.

4.1 Design of the System

The propose system will classify Myanmar language online news documents into one of the predefined category such as Politic, Business, Entertainments, Sport. The system contains two phases in the system: training and testing.

For training phases, online news documents are collected manually and preprocessed .Then, they are stored as a labeled training data set. And, using TF-IDF to calculate how important word is for a documents or corpus and store the weight of each word and store for the later use.

In testing stage, the system can accept the news document that contains text data. After accepting the input, classification system performs the preprocessing step. The preprocessing consists of two steps including word segmentation and stop word removal. The first step in preprocessing is word segmentation, in this step each sentence in the text document is segmented into word. After sentences are segmented, some commonly occurring word need to be removed from the list of words such that ‘လာမည်’,’ နေပြီ’,’ ပါသည့်’ etc. This process is known as stopword removing .The system maintains a list of word and it is checked against with the stopword list to remove the word. After stopword removal is performed, there are still many words exit in the system due to the nature of text document. Genetic Algorithm is used to classify documents which will calculate the best possible answer among these words.

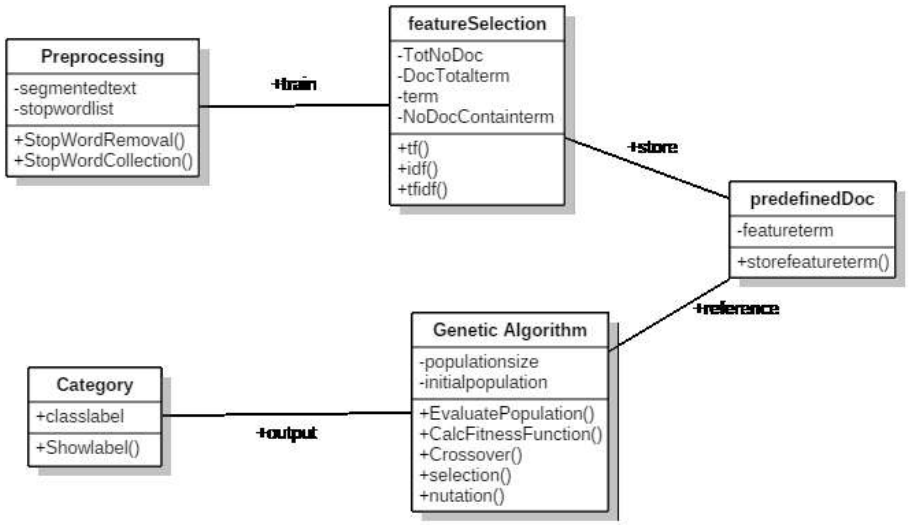

Figure 4.1: Class diagram for Myanmar Text Classifier using Genetic Algorithm

4.2 Implementation of the System

This section describes how to implement text classification system for Myanmar news articles starting from the stage of data collection to measurement of performance of the system.

4.2.1 Data Collection

For the implementation of this classification technique, data resources are collected from Eleven Media group, 7 days Daily, Irrawaddy Burmese, VOA Burmese, BBC Burmese, Sport Myanmar, popular news journal. This thesis will classify based on 4 categories for the research and they are listed as below.

1. Politic
2. Business
3. Sport
4. Entertainment

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

Table 4.1: Document Collection for Training and Test Data.

The collected documents include news in Myanmar Language which is composed of pure text data and speech transcriptions.

4.2.2 Preparation for Training Dataset

First of all, the system uses the Myanmar Unicode encoding system. So, it is important to make sure that all the collected documents are in Myanmar 3 Unicode font. Then, collected documents are needed to preprocess for clean dataset. As, two stages of the preprocessing stage, word segmentation and stopword removing

(a) Word segmentation

The proposed system will accept the input as the already segmented words. In this case, the word segmentation is done by the outside of the system and segmented according to the segmentation rule manually. And segmentation rule has been described in chapter 3.

(b) Stopword Removal

Then, the system will remove stop words according to stop words list. During the segmentation stage, the documents are used not only for segmentation but also for the storage of the stopword list. As the segmentation is done manually, the only way to store the stopword list is to done by manually. In stopword segmentation, punctuations, other unnecessary characters are also removed because these words do not help in deciding the category of the documents. More and more stop words are needed to be collected to produce better feature words and to improve performance of the system. Figure 4.2 shows a list of simple collected stop words.

တွေ	ရထားပြီး	ဖြစ်ဦး
ည	နေရတယ်	မလဲ
ကြတော့	နိုင်ဘူး	ဒါပေမဲ့လည်း
ရာသို့	ပိုကြုံ	ချင်တော့
လိုသည်	ပေါ့နော်	ပါပြီး
ဖြစ်နေကြောင်း	ဟယ်	ဖြစ်ဖို့
နေပြီ	လို့ပြော	ကြမှာလဲ
ရှိနေမည်	ပါတဲ့	ဖြစ်မလဲ
ထားခဲ့	ကိုယ်လည်း	အတူ
အဲဒီမှာ	အဲဒါ	သူ့ကို
မြင်ကြ	က	မို့လို့
လိုခြင်း	တော့	ဖြစ်သွားတော့
သကဲ့	ပြောခဲ့သည်	ကပေါ့
ခဲ့ရာမှ	ကနေပြီး	သလောက်
လာကြောင်း	ဒီလိုမှ	ချင်တာ
ရှေ့	ရခဲ့ကာ	တာကြောင့်
မနက်ဖြန်	ရှိလို့	မြင်တယ်
နေရကြောင်း	ခါနီးမှာ	ကျတယ်
နိုင်သည်	တွေလိုပဲ	ထိုစဉ်
ယခုထက်	ဖြစ်လာကြောင်း	လာသော်လည်း
ဒါပါဆိုပေါ့	ဖြစ်နေခြင်း	တာပေါ့
	ဖြစ်ရသည်	ရှိလာမည်
	ရှိသည့်	ဗျ

Table 4.2 Stop words list

4.2.3 User Interfaces of the System

The following pictures are the screen shot of user interfaces of the applied system "Myanmar Text Classifier Using Genetic Algorithm". The user interface of the system includes two parts namely training phase and classification phase. Figure 4.3 show Main Page of the system.

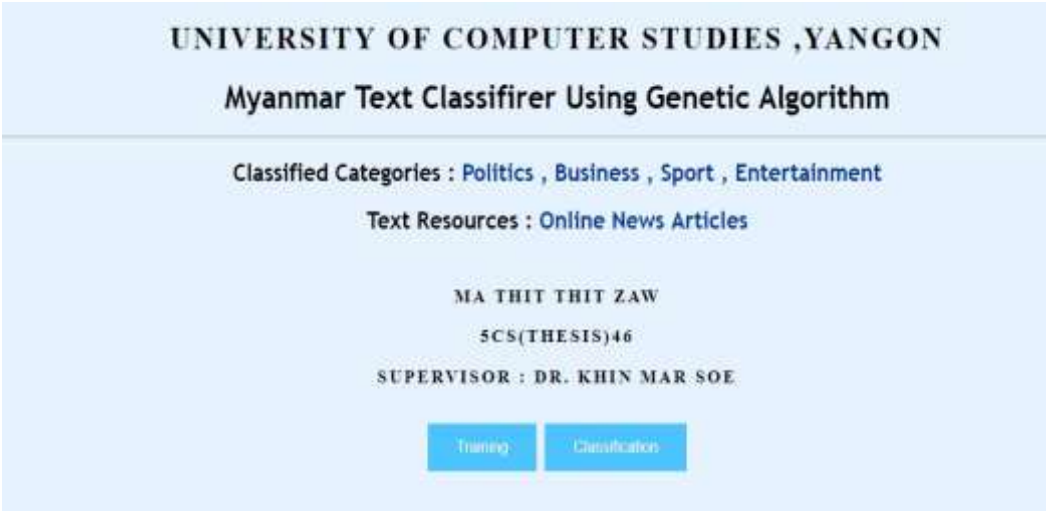

Figure 4.2 Main Page of the System

In training phase, the system needs the label of the documents and documents to train in order to collect the feature according to their categories. Figure 4.4 show the training page of the system.

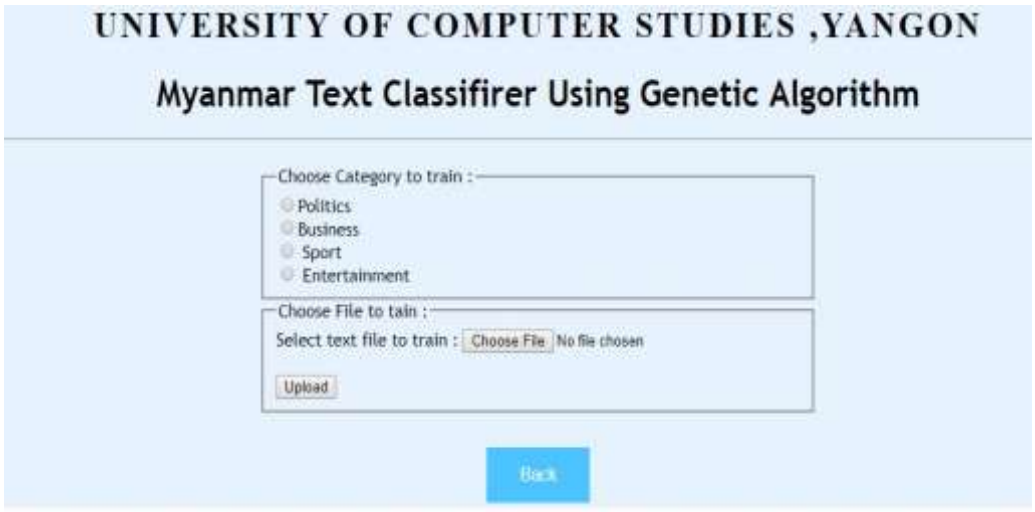

Figure 4.3 Training page of the system

In training process, the page will display category label to train and file selector to select text file to train.

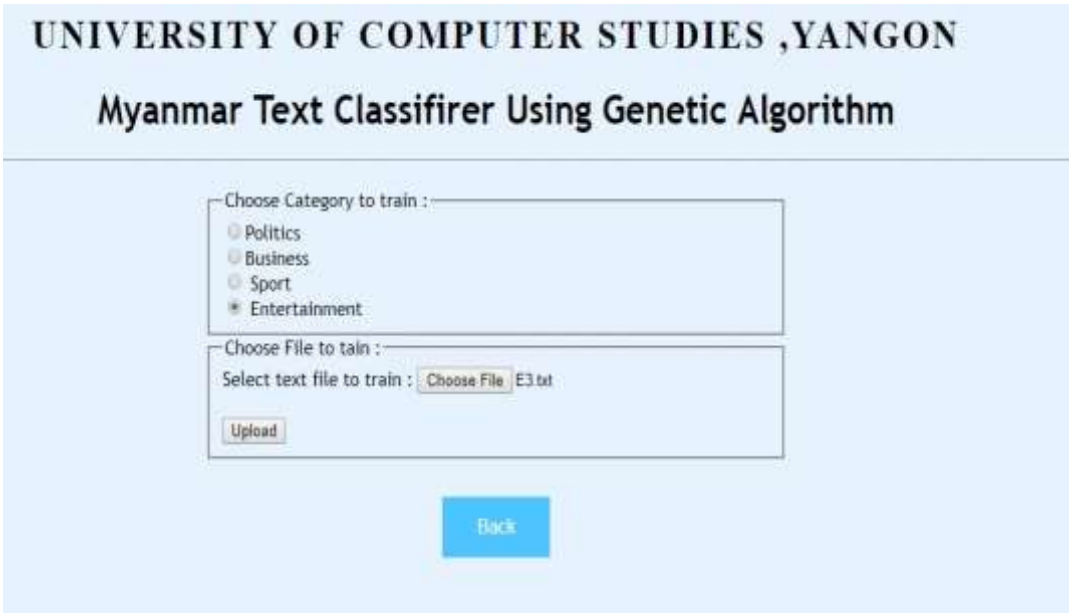

Figure 4.4 Training Document

The user will need to choose label and the document to be trained by using Upload button. After uploading the text document and category to train, the system will use the TF-IDF algorithm to collect the related feature and store the related feature based on the user input category .Back button will allow the user to go back to the Home page of the system.

Figure 4.5 Page View after Training Documents.

After uploading the document, it will lead to new page and show the content of the document and show the message about the category label. Save button

will take to go back to training page where the user can choose label and upload document. Figure 4.6 show the next page of the training phase and text area show the content of the training document with the message which show the collected feature are store under the provided category.

In classification phase, the documents of unknown label are used to classify.

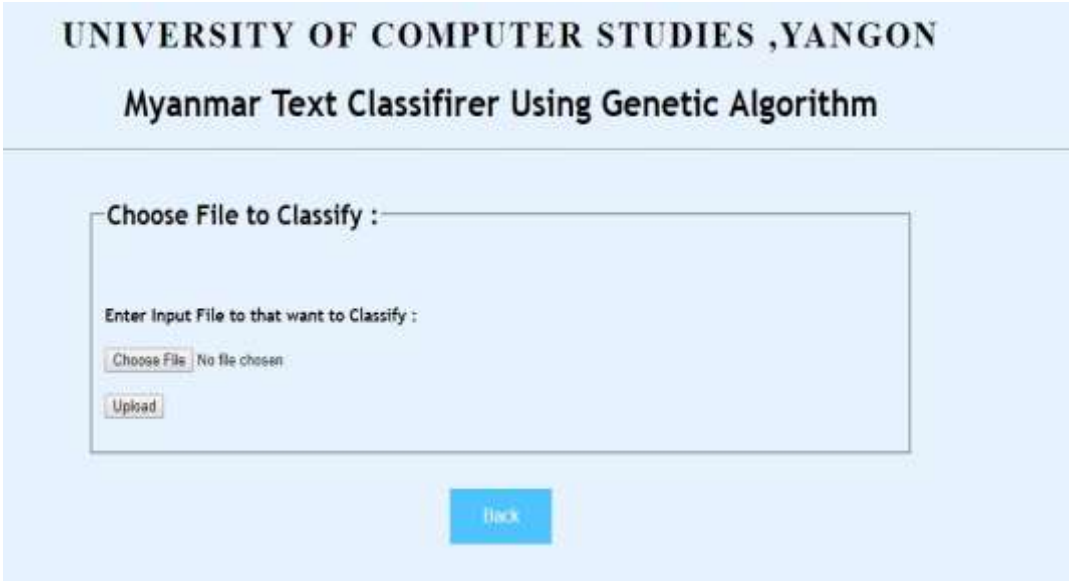

Figure 4.6 Classification Page

Figure 4.7 show the classification page of the applied system. The page contains the file selector to choose text file to classify and Back button will go back to the Main page of the applied system.

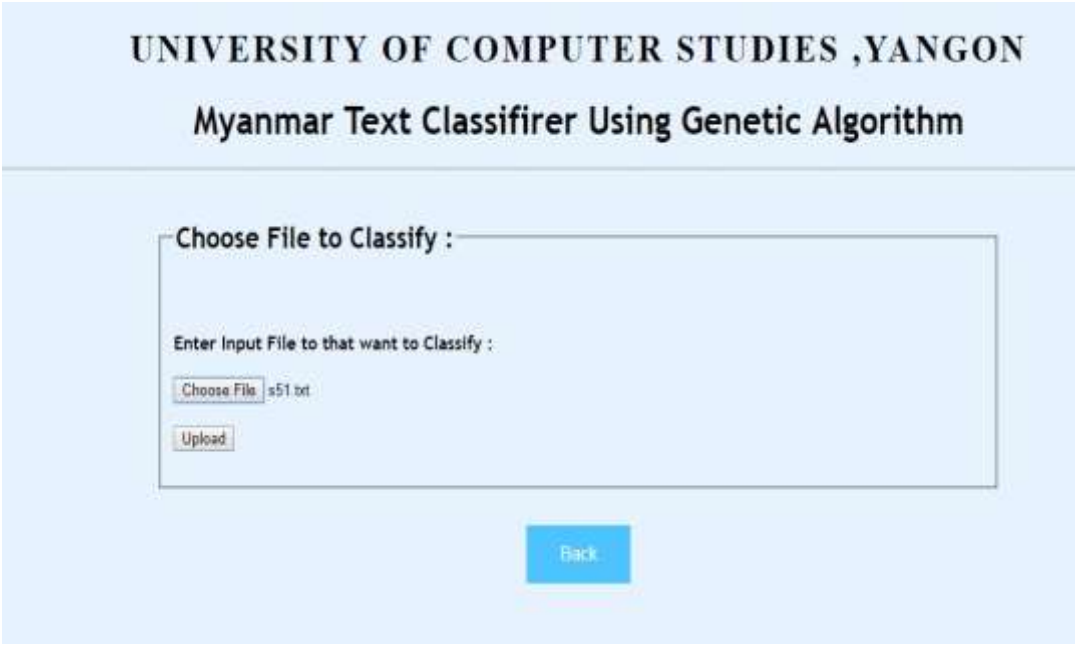

Figure 4.7 Classifying Document

In classification stage, the system needs to provide text document to train and Back button can take back to the main page.

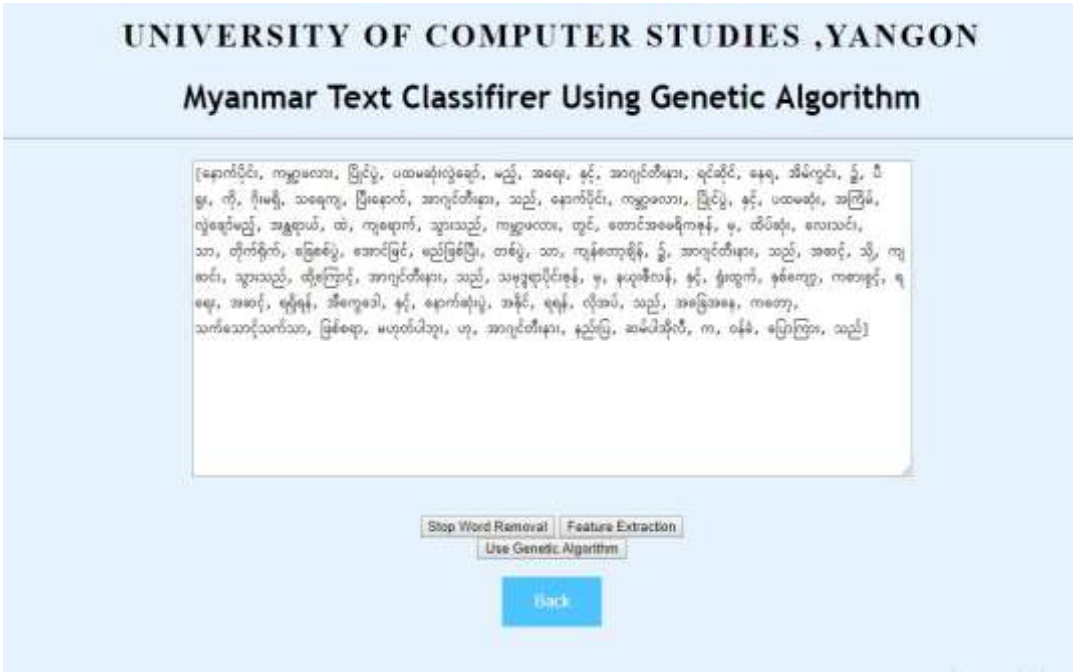

Figure 4.8 Initial Stage of the classification.

The above figure 4.9 shows the content of the chosen test file in the text box. The page contains the three buttons Stop Word Removal, Feature Extraction and Use Genetic Algorithm. The Stop Word Removal Button will remove the stop words from the content of the document. Feature Extraction Button will extract only

features according to the collected feature and the Use Genetic Algorithm buttons use the Genetic Algorithm to classify the input text file.

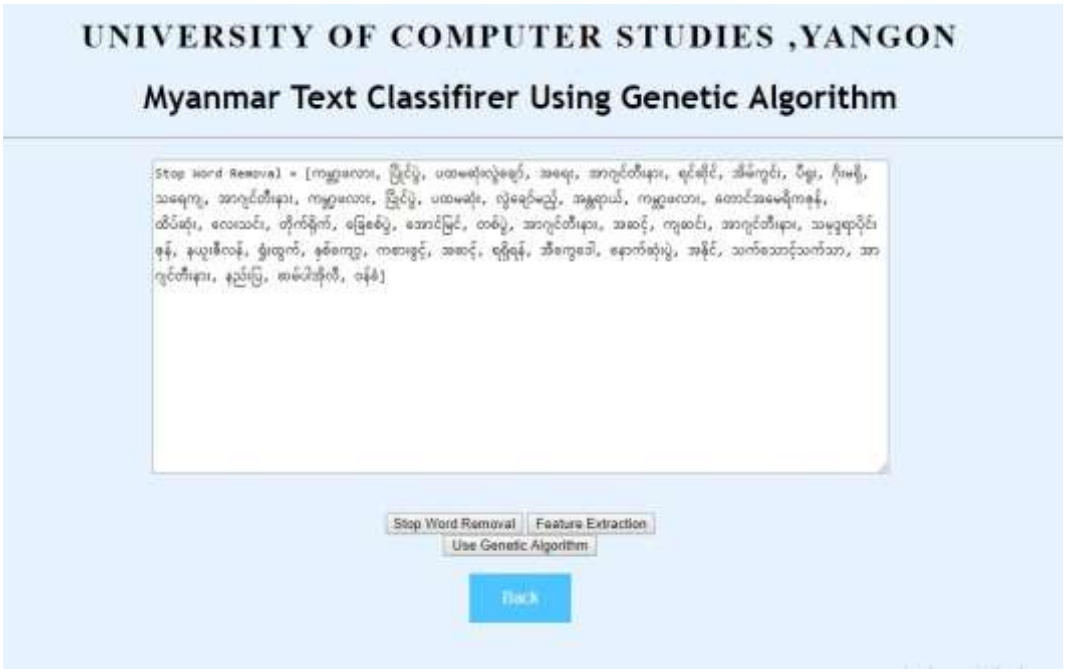

Figure 4.9 Stop Word Removal Stage

After clicking the stop word removal button, the system will remove stop words from input document according to the term collected in the stop words.txt and show it in the text box. Figure 4.10 show content of the text document which already removed stop words.

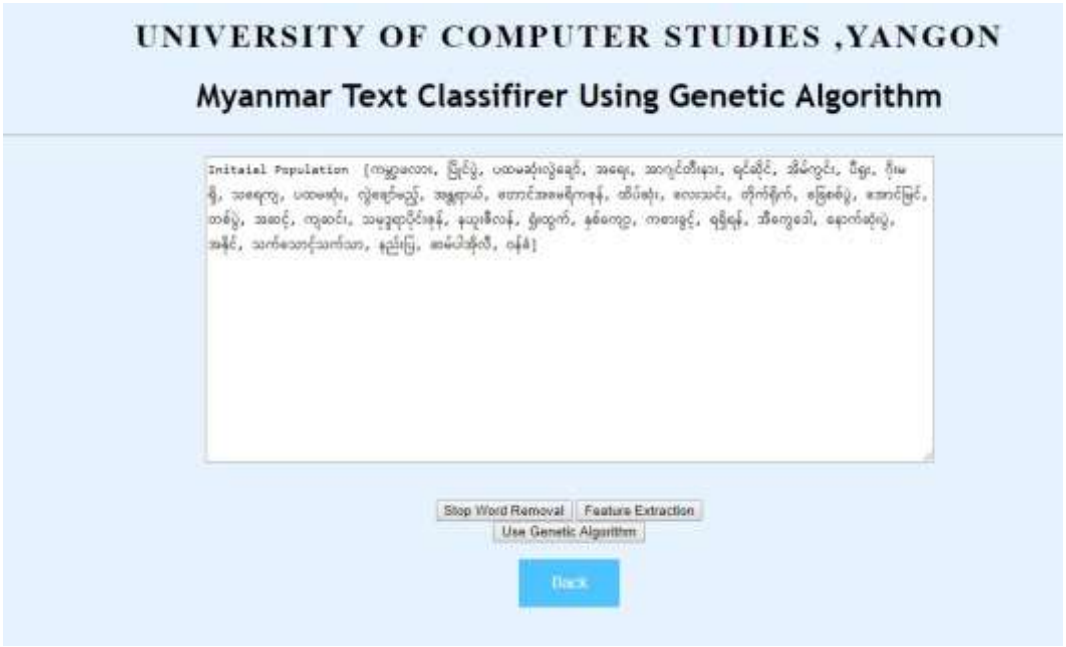

Figure 4.10 Feature Extraction Stage

Feature extraction button will extract the feature form the input documents after stop words removal. Figure 4.11 show the feature of the input text file.

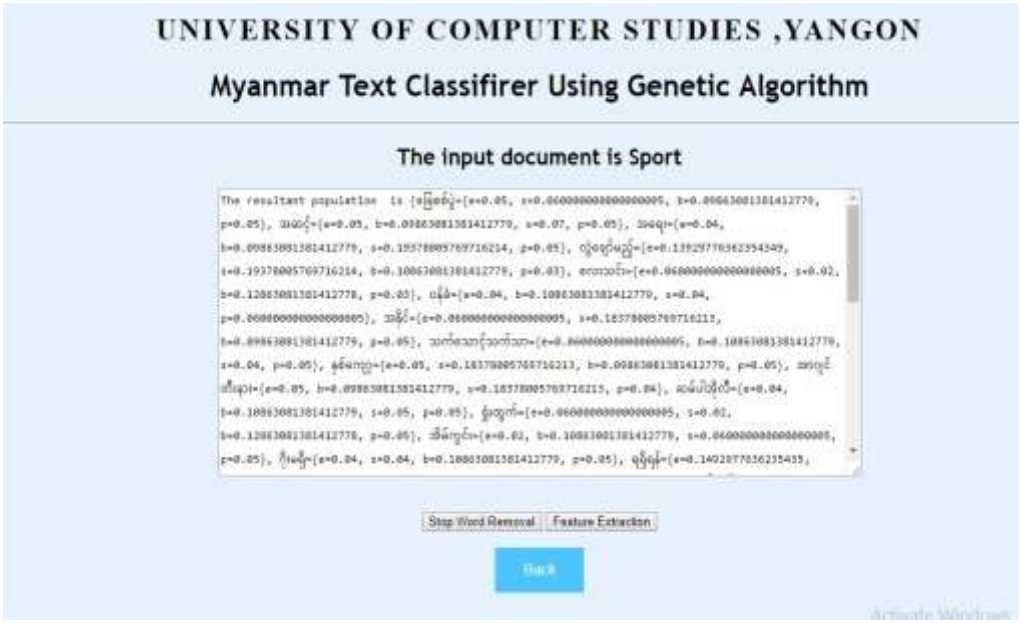

Figure 4.11 Category of the input document

Use Genetic button will display the predicted category label and text area shows final population of the document by using genetic algorithm to classify and Back button will take back to the classification page. Figure 4.12 show the category of the provided document.

4.3 Experimental Result

In this system, a document is assigned to only one category. Precision, recall and F-measure are used as performance measures for the test set. In performance measuring, 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. 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: $$\frac{Number\ of\ correctlyclassified\ documents\ \text{to}\ acategory}{Total\ number\ of\ documentslabeled\ bythe\ systemasthat\ category}$$ Eq [4.1]

Recall: $\frac{Number\ of\ correctly\ classified\ documents\ \text{to}\ a category}{the\ number\ of\ documents\ of\ that\ category\ \text{in}\ training\ data}$ Eq [4.2]

F1 (recall, precision): $2.\frac{(precision.recall)}{(precision+recall)}$ Eq [4.3]

Accuracy	**politics**	**Business**	**entertainment**	**Sport**
Doc 100	0.44(44%)	0.6(60%)	0.76(76%)	0.6923(69%)
Doc 150	0.6(60%)	0.76(76%)	0.76(76%)	0.577(57%)
Doc 200	0.84(84%)	0.84(84%)	0.8077(87%)	0.80777(87%)

Table 4.3 Accuracy Measurement of the System

CHAPTER 5
CONCLUSION

This thesis describes the automatic text classification system for Myanmar news articles using Genetic algorithm for classification purposes. Text mining and text classification techniques are discussed in detail in Chapter 2. Theoretical details of the system are described in Chapter 3 and implementation details are illustrated in Chapter 4. Since this system applies semi-supervised classification method for text classification system in Myanmar language, it has some weaknesses that need to be considered to solve in future.

In order to classify Myanmar news articles, a text classifier based on Genetic algorithm is implemented and TF_IDF feature selection algorithm is applied to measure how important of words to the certain corpus for classification task. News from Myanmar online media websites are collected manually and labeled. Then, these news documents are preprocessed and stored as a training corpus. The system has been tested using real-world news data collected from reliable Myanmar media websites. The accuracy of the system is measured by Precision, Recall and F-l measure methods. It has been found out that high-quality feature words can improve the performance of the system. According to the nature of Genetic algorithm and testing results, it can be concluded by saying that if the training dataset contain the more high-quality data, the accuracy of the system will also increase.

5.1 Advantages and Limitations of the System

The main advantage of Genetic Algorithm is very easy to understand and don't need to provide the prior knowledge about the problem. Although it requires time to compute for calculation, to know irrelevant features and to apply all the possible way for the better solution. On the other hand, Genetic algorithm classifier needs a lot of training data to obtain good results. The more training data, the better performance of the system because the quality of features depends on the segmentation of words. So, when more words can be added to word segmentation application, the better feature words can be obtained. Some failures in the system are caused by the amount of training data and word segmentation problem. Another limitation is that the system works well only on well-defined categories and no

bounded category should be considered for invalid data. The proposed system can be improved by adding more data to the training documents to increase the accuracy of the classification task.

5.2 Application Areas

Text classification can be applied in search engines, recommender systems and other information retrieval applications. The system can be applied in many real world applications in Myanmar Language.

Publications

[1] Thit Thit Zaw, Khin Mar Soe, "*Myanmar Text Classifier Using Genetic Algorithm*", to be published in the Proceedings of the 9th Conference on Parallel and Soft Computing (PSC 2017), Yangon, Myanmar, 2018.

REFERENCES

[1] Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski, Michael W. Mahoney.," Feature Selection Methods for Text Classification", Research Track Paper, in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 230-239 , August 12 - 15, 2007, San Jose, California, USA.

[2] Aye Hnin Khine ,"Automatic Myanmar News Classification using Naïve Bayes Classifier", Thesis Book, University of Computer Studies, Yangon (UCSY), August 2017, Yangon.

[3] Charu C.Aggarwal and Cheng Xiang Zhai, "Mining Text Data", Book, ISBN: 1461432227 9781461432227, June 2012.

[4] Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita, National Institute of Information and Communications Technology," Word Segmentation for Burmese (Myanmar)",Journal, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Volume 15 Issue 4, June 2016 ,Article No. 22

[5] David E. Goldberg., "Genetic Algorithms in Search, Optimization, and Machine learning".Book, ISBN: 0201157675, Boston, 1989.

[6] Divya P, G.S. Nanda Kumar," Study on Feature Selection Methods for Text Mining", International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)Vol. 2, Issue 1, January 2015.

[7] Dr.Khin Mar Soe, Dr. Khin Thandar Nwet, Aye Hnin Khine,"Automatic Myanmar News Classification System", conference paper,15th International Conference on Computer Applications (ICCA 2017), Yangon, February 2017.

[8] Dr. S. Vijayarani , Ms. J. Ilamathi , Ms. Nithya ,” Preprocessing Techniques for Text Mining - An Overview”, ISSN:2249-5789,International Journal of Computer Science & Communication Networks,Vol 5(1),7-16, (February-March) 2015.

[9] Gurpreet S. Lehal and Vishal Gupta,” A Survey of Text Mining Techniques and Applications”, Journal, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[10] Michael Steinbach, George Karypis, Vipin Kuar,”A Comparison of Document Clustering Techniques”, Technical Report, in proceeding of the KDD Workshop on Text Mining,2000.

[11] Myanmar Thuddar { mrûn-ma þûd~da}, Volume 1, Module 1, by Myanmar Language Commission (MLC), Ministry of Education, Government of the Union of Myanmar (in Burmese-Myanmar) , Date of publication: around 1986.

[12] Radha Guha , “Exploring the Field of Text Mining”, International Journal of Computer Applications (0975 – 8887) Volume 177 – No.4, November 2017.

[13] RUIZ, Miguel E.; SRINIVASAN, Padmini. “Automatic Text Categorization Using Neural Networks”, Advances in Classification Research Online, [S.l.], p.58-68, Nov.1997.ISSN2324 9773. doi:http://dx.doi.org/10.7152/acro.v8I1.12728.

[14] S. M. Kamruzzaman , Farhana Haider And Ahmed Ryadh Hasan ,” Text Classification Using Data Mining”, Journal, International Conference on Information and Communication Technology in Management (ICTM-2005), Multimedia University, Malaysia, May 2005.

[15] S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil," Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution", Academic Journal ,World Academy of Science, Engineering and Technology International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering Vol:2, No:1, 2008.

[16] Sung-Sam Hong, Wanhee Lee, and Myung-Mook Han ," The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification ",International Journal of Advances in Soft Computing and its Applications, Vol. 7, No. 1, March 2015, ISSN 2074-8523.

[17] Sanasam, R., Murthy, H. & Gonsalves, T," Feature Selection for Text Classification Based on Gini Coefficient of Inequality", Proceedings of the Fourth International Workshop on Feature Selection in Data Mining, in PMLR, January 2010.

[18] Taiwo Oladipupo Ayodele (February 1st 2010)." Types of Machine Learning Algorithms", New Advances in Machine Learning Yagang Zhang, Intech Open,DOI:10.5772/9385.Availablefrom:https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms.

[19] Vijini Mallawaarachchi (2017, July 7),"Introduction to Genetic Algorithms—Including Example Code", https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3.

[20] Win Win Thant,Tin Myat Htwe and Ni Lar Thein ,"Syntactic Analysis of Myanmar Language", Proceedings of International Conference on Computer Applications (ICCA 2011), Yangon, Myanmar, May 5-6, 2011.

[21] Yiming Yang, Jan O.Pedersen .,"A Comparative Study on Feature Selection in text Categorization", conference paper, ICML '97 Proceedings of the

Fourteenth International Conference on Machine Learning, Pages 412-420,July 08 - 12, 1997.

[22] Young Joong Ko, “A Study of Term Weighting Schemes Using Class Information for Text Classification”, in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). ACM, New York, NY, USA, 1029-1030, August 12 - 16, 2012, DOI: https://doi.org/10.1145/2348283.2348453, ISBN: 978-1-4503-1472-5.

[23] Zhu Wei Dong, Feng Jing Yu, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, and Lin Yong Min, College of Economics And Management, Hebei Polytechnic University, Tangshan 063009, China,” Using Gini-Index for Feature Selection in Text Categorization”, 3rd International Conference on Information, Business and Education Technology (ICIBET 2014).