Feature Selection for Categorization of Online News Articles in Myanmar Language

Myat Sapal Phyu, Win Win Thant, Thet Thet Zin
University of Information Technology, Yangon, Myanmar
myatsapalphyu@uit.edu.mm, winwinthant@uit.edu.mm, thetthetzin@uit.edu.mm

Abstract

In text mining, the feature selection plays an important role to reduce the high dimensionality of feature space. It can improve the accuracy of the document clustering process and help to avoid overfitting problem. Nowadays, the enormous amount of news article documents is widely available on the internet due to the rapid development of the web. Consequently, there is an urgent need to extract useful content from overloaded information. The categorization of online text documents is crucial to avoid information overload and it can help readers to find rapidly their interesting topic. The problem arises for text categorization is the large number of features space. This study has two phases, documents preprocessing and feature selection. Document preprocessing contains documents collection, syllable segmentation, word segmentation, removing stop words for extracting features from the collection of Myanmar online news documents including sport, health, crime etc. In this study, TF-IDF weighting method is adapted for feature selection. The experimental result shows the adapted TF-IDF method has higher performance than based TF-IDF method.

Keywords- Feature Selection, TF-IDF, Syllable Segmentation, Word Segmentation, Myanmar Online News

1. Introduction

In recent year, the rapid growth of using the internet leads to the information overload. People get a lot of information through the internet every day and waste much time to select their interesting information. Consequently, there is an urgent need to extract useful content from the enormous amount of information quickly and effectively. The categorization of news events is an important research area in text mining. It can enable the aggregation of news stories by topic and provide the basis for news recommender systems, a subclass of information filtering system. It can help readers to get the brief information about news documents before they read it. It is the way to automatically categorize news documents into predefined categories such as sports, education, and technology etc. The main difficulty in text categorization is the high dimensionality of feature space. Ordinarily, the large number of features presents in the collection of documents and a few are informative. Accordingly, some important features are needed to select to reduce the dimensionality of feature space because it contributes directly to the accuracy of the document clustering. This study focuses on feature selection for categorization of Myanmar online news articles.

In this study, some preprocessing steps are needed to perform before extraction and categorization of news articles. The preprocessing steps contain listing stop words, syllable segmentation and word segmentation. Word segmentation is performed for extracting features (words) from collection of text documents.

Feature dimension reduction is an important part in text categorization. This study analyzes the TF-IDF method and makes adaption based on this method in order to get high accuracy for feature selection. In information retrieval, the TF-IDF is a well-known method to evaluate how important is a word in a document. It is a very interesting way to convert the textual representation of information into a term vector model. It is an algebraic model for representing text documents as vectors of identifiers.

The first step in converting the document into a vector space is to create a dictionary of terms (words) by selecting all terms from the document and transform it to a dimension in the vector space. As the main task is to select important features from documents, the stop words are ignored because they are not helpful to categorize the text documents. So, stop words list in Myanmar language is created by analyzing Myanmar online news documents to remove unnecessary features.

In order to extract features from collection of documents, it needs some preprocessing steps. In preprocessing step, the syllable segmentation and word segmentation are considered in order to specify each separate word as one feature. After extracting features, important features are selected by adapted TF-IDF method and compare the performance with existing TF-IDF method.

The rest of this paper is as follows, section 2 describes the related research that was published in the area of Myanmar word segmentation, syllable

segmentation and feature selection methods. Section 3 discusses the overview of feature selection process including collecting text document from Myanmar daily news websites [12] [15], preprocessing tasks, feature selection by TF-IDF method and its adaptation. In section 4, the nature of data set is discussed. The detail steps of preprocessing steps are explained in section 5. In section 6, feature selection by TF-IDF [10] method is presented. According to the testing result of baseline method, it is adapted in TF-IDF (Adaptive Method) in order to get better solution and discuss in section 7. Experimental results are discussed in section 8. According to the experiments, some problems are pointed in section 9. The last section concludes and discusses the future works.

2. Related Work

Document pre-processing and feature selection approaches are useful for text categorization process. Myanmar word segmentation and syllable segmentation play an important role in document pre-processing task. Many researchers did in Myanmar word segmentation [2] and syllable segmentation by using different methods [6] [7] [11] [2]. The feature selection approaches are most important research area in text mining and implemented by different methods [3] [4] [5].

Manually constructed context free grammar (CFG) is presented in [6] to describe the Myanmar Syllable Structure to identify Myanmar syllables. The syllable segmentation algorithm that can slice all of the input text string is developed in [7]. The input text strings are converted into equivalent sequence of category form and compare the converted character sequence with the syllable rule table to determine syllable boundaries. A syllable segmentation tool is developed for Myanmar text encoded with Unicode in [11].

The two steps method for syllable segmentation and syllable merging are proposed in [2]. Syllable boundaries are determined by the proposed six syllable segmentation rules and dictionary-based statistical approach is used to perform syllable merging.

In [3], terms are extracted from the documents by using term selection approaches tf-idf, tf-df and tf2 based on their minimum threshold value. This approach is intended to reduce the attributes and find the effective term selection method using WordNet.

In [4], Support Vector Machine is applied to classify Bangla document and TF-IDF is used for feature selection.

A new weighting method named TF-IDF-CF is proposed in [5] based on TF-IDF. It introduced a new parameter class frequency, which calculates the term frequency in documents within one class.

The main aim of this study is to adapt TF-IDF (Baseline Method) by analyzing, testing the baseline method with the input data set, Myanmar text documents collected from Myanmar online news websites [12] [15]. The experimental result shows the better performance of the TF-IDF (Adaptive Method) than baseline method. In the future, various feature selection methods will be tested and implemented with more data set in order to find better solution and intend to find high informative features for categorizing Myanmar online news documents in the future. It is intended to support the Myanmar online news categorization process to overcome the difficulties of Myanmar news readers to get their desired news rapidly.

3. Overview of Feature Selection Process

Figure 1. shows the overview of the feature selection process for categorizing Myanmar online news articles. Firstly, text documents are collected from social news websites. Then, the features are extracted from input text documents.

In order to extract features, input text documents are segmented into syllables and merged these syllables into meaningful word by matching Myanmar words dictionary [13] and then unnecessary words including city, date time words, number, non-Myanmar character and so on are removed. After that, features are selected by TF-IDF (Baseline Method) [10] and adapted the baseline method in order to get high accuracy. Then the selected features are collected into lexicon for the future categorization process.

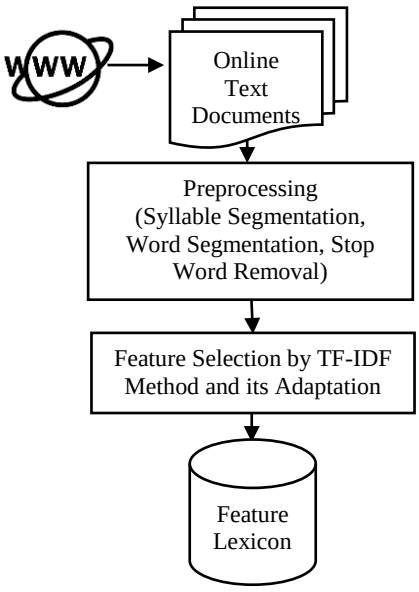

Figure 1. Feature Selection System Design

4. Data Set

Text data are collected from a Myanmar daily news website [12] [15] and extracted text by online text extractor [9]. Then, extracted text are converted into Unicode format by Zawgyi-One to Unicode converter [14]. Each news article is saved as text document (.txt) and used as input data set for feature selection process. Each news article generally contains about 10 sentences. In this study, 383 news articles for crime category, 320 news articles for sport category and 283 news articles for health category are collected as text documents and more articles for updated news will be collected in the future. Table 1. shows the data set on three different news categories. 16,381 features from 383 crime news, 12,273 features from 283 health news and 14,801 features are extracted from sport news.

Table 1. Data Set on Three Online News

No.	Category	Number of Documents	Number of Extracted Features
1.	Crime	383	16,381
2.	Sport	320	14,801
3.	Health	283	12,273

5. Preprocessing

Before the extraction of features from news documents, the following preprocessing steps are required to extract the features:

5.1. Syllable Segmentation

In order to extract the news features, it is needed to determine the word boundaries. The syllable boundary is determined as the preprocessing task for word segmentation. In this study, syllables are segmented by syllable segmentation method that is implemented by regular expression pattern [11]. The pattern is based on encoding order of Myanmar syllable.

According to the encoding order of Myanmar syllable [8], most of the Myanmar syllables start with consonant and the starting character of each syllable can also be determined as end of preceding syllable. So, the syllable boundary is determined by checking the consonant and marks with syllable boundary marker in front of the consonant except the consonant after a subscript character (“္”), the consonant that is followed by a-that character (“်”) or subscript character (“္”) and standalone syllable such as (“ဣ, ဤ, ဥ, ဦ, ဧ, ြ, ျော်, ဿ, ၌, ၍, ၏”). This approach can reduce time and space complexity and outperform than other syllable segmentation methods.

5.2. Word Segmentation

After segmenting each syllable, the segmented syllables are merged to form a word. The input text is segmented into each individual syllable. And then, segmented syllable are merged to become meaningful word by dictionary based maximum syllable longest matching approach using Myanmar words lists [13]. Myanmar words list contains 41482 words and some missing words in dictionary that are mostly used in sport, crime and health online news are added to this list by analyzing 383 crime documents, 320 sport documents and 283 health.

For instance, the popular football terms "ဦးဆောင်ဂိုး, ချေပဂိုး, စပေးဘောလုံး", the name of famous professional footballer and the coach such as "စီရော်နယ်ဒို, အာစင်ဝင်းဂါး" and the most popular virus and diseases in medical field such as "ရာသီတုတ်ကွေး, လူတုတ်ကွေး, ဇီကာပိုင်းရပ်စ်" are considered to be added to the Myanmar words lists. The words related with sport domain are mostly added to the dictionary because most of the sport news are international news and not contain in dictionary. Currently, 41767 words are presented in the Myanmar words list.

5.3. Removing Stop Words

Stop words are the set of commonly used words in any language. They are removed from feature space to reduce the noise and to enhance the computational efficiency of categorization.

In this study, stop words are collected by analyzing Myanmar online news. Most of the news contains the location and time information that are not important terms for categorizing news documents. After analyzing news documents, location, date, time words and the most commonly used prepositions, inflections; conjunctions are collected as stop words. Moreover, punctuation marks (eg., "။ ."), white spaces and other symbols (eg., "-/()[]{}"), non-Myanmar text (eg; A to Z), numerical text (eg., 0 to 9 ๐to ๙) are removed. In this study, 608 stop words are collected and more stop words will be added in the future. Table 2 shows the sample of stop words list.

Table 2. Sample of Stop Words List

ဧပြီ၂၄ရက်၊ ဇူလိုင်၊ ဩဂုတ်၊ နာရီ၊ မိနစ်၊ စက္ကန့်၊ နေ့လည်၊ နှစ်တော်၊ ပဉ္စသို့၊ တပို့တွဲ
တိုင်းဒေသကြီး၂မြို့နယ်၊ မြို့သစ်၊ ကျေးရွာအုပ်စု၊ ဘိုကလေး၊ ခနဖျူး ဒေးဒရဲ၊ မ်ဂလာဒုံ။
နောက်ထပ်၊ တခြား၊ နောက်တစ်ခု၊ နောက်တစ်ချက်၊ နောက်တစ်ယောက်။
သည့်အပြင်၊ ထို့ပြင်၊ ထို့အပြင်၊ ဒါ့အပြင်၊ နောက်ပိုင်း၂

6. Feature Selection by TF-IDF (Baseline Method)

Features are selected and tested by adapted TF-IDF and original TF-IDF method, short for Term Frequency–Inverse Document Frequency that is intended to reflect how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document (one domain), but is offset by the frequency of the word in the corpus (all domains), which helps to avoid multi label problem.

$$TF - IDF = TF(t,d) * IDF(t,D) \text{ [10]}$$

$$= \frac{Freq(t,d)}{|D|} * log \frac{N(D)}{Freq(t)} \quad (1)$$

Freq (t, d) - the frequency of term t in one domain d (among sport or crime or health)
|D| - total number of term in the domain

IDF - Inverse Document Frequency

N (D)- number of domains

Freq (t) - number of domains that contain term t

TF-IDF method is tested with 16,381feature (terms) extracted from 383 crime news documents, 12,273 features (terms) extracted from 283 health news documents and 14,801 features (terms) extracted from 320 sport news documents from Myanmar news media sites [12] [15]. TF value is calculated within one domain. and IDF value is calculated across the domains. TF-IDF score of each feature is calculated and ranked by TF-IDF score and top 100 features will be selected as important feature for each domain. According to the some findings on TF-IDF (Baseline Method), it is adapted in TF-IDF (Adaptive Method) and the details are briefly explained in section 7.

7. Feature Selection by TF-IDF (Adaptive Method)

The concept of TF-IDF [10] is that TF-IDF score is highest when TF value is high (high frequency of term in one domain) and IDF value is high (does not contain frequently in all domains).

In this study, TF-IDF (Baseline Method) is adapted mainly for two kinds of features. Firstly, the features that contain in few domains (high IDF value) but it does not have much frequency in each domain (low TF value) have low TF-IDF scores even though they may be high informative than other features that have high term frequency and low IDF value. To solve this problem, if the value of term frequency is higher, its value may raise than other features that contain in many documents. If we calculate the term frequency value with logarithm in both numerator and divisor as TF=log(Freq(t,d))/log(|D|) , the quotient will be little higher than original TF method.

For instance, suppose that we have to consider four domains sport, education, health and sport (N(D)=4) and each domain has 500 extracted features (|D|=500). The feature "*Champion*" contains 10 times in sport domain ($Freq(t,d) = 10$) but not contains in other (Freq(t)=1). Then, the feature "*Agreement*" contains 50 times in sport domain ($Freq(t,d) = 50$)and it also contains in other two domains (Freq(t)=3). TF score of the feature " *Champion*" is 0.04, IDF score is 0.6021 and TF-IDF score of the feature "*Champion*" in sport domain is 0.012042.

TF score of the feature "*Agreement*" is 0.2, IDF score is 0.1249 and TF-IDF score of the feature "*Agreement*" in sport domain is 0.01249. According to the TF-IDF results, the feature "*Agreement*" that contains in many domains with high term frequency has higher TF-IDF score than the feature "*Champion*" that contains in only one domain that has low term frequency. The results are depicted in Table 3.

Table 3. Comparison TF-IDF Scores of Two Features by TF-IDF (Baseline Method)

Feature	TF	IDF	TF-IDF
Champion	0.04	0.6021	0.012042
Agreement	0.2	0.1249	0.012490

The purpose of adapting equation is to raise the value term frequency value for the features that only contain in one domain than the features that contain in many domains. By adjusting these values, the accuracy of selected features is higher than baseline method according to the experiments. Table 4. shows the TF-IDF sore of the feature "*Champion*" and "*Agreement*" by adapted TF-IDF method. the feature "*Champion*" get higher TF-IDF score than the feature "*Agreement*". In these tables data are simplified for illustrative purpose.

Table 4. Comparison TF-IDF Scores of Two Features by TF-IDF (Adaptive Method)

Feature	TF	IDF	TF-IDF
Champion	0.3705	0.7021	0.2601281
Agreement	0.6295	0.2249	0.1415746

Secondly, in case of the features that have IDF value zero, IDF value is added 0.1 for smoothing as IDF=log(N(D)/Freq(t))+0.1. Suppose that, we have to consider for two domains, sport and crime. For instance,

the term "Football" is frequently found in sport news. So, term frequency of "*Football*" in sport domain may be high. The term "Football" is also found in crime news as "*Sixteen-year-old football player is killed instantly after the 10-foot log struck him on Thursday*". Although the term "*Football*" is found in crime news, it is a condition that is not occurring very often that term frequency of "*Football*" in crime domain may be low. If we use as log(N(D)/Freq(t)), the score of TF-IDF value for both domain will be zero because log(2/2)=log(1)=0. If we add 0.01, TF-IDF score will be higher that has higher term frequency value (sport domain).

In brief, the adapted TF-IDF method is to justify mainly for features that have very low term frequency with high IDF value and justification of TF-IDF value that has IDF value zero. Equation 2 shows the adapted TF-IDF method.

$$TF - IDF = \frac{\log(Freq(t,d))}{\log(|D|)} * \log\left(\frac{N(D)}{Freq(t)}\right) + 0.1 \quad (2)$$

8. Experimental Results and Discussion

In this study, 16,381 crime features, 14,801 sport features and 12,273 health features are extracted from 320 sport documents, 383 crime documents, and 283 health documents. Then features are selected by TF-IDF model and its adaptation model and then the performance is experimentally evaluated. Table 5. (a), (b), (c) show the top 8 terms for each category including sport, health and crime, ranked by TF-IDF score. In these tables, data are simplified for illustrative purposes. In reality, top 100 features are selected as important keywords for each category.

Table 5. (a) Sport Terms

Sport Category	
Term	**TF-IDF**
ပို့ဂ််ပွဲ	0.61569
ယှဉ်ပို့ဂ်	0.59391
ကစားသမား	0.58658
ပရိသတ်	0.52659
အိုလံပစ်	0.51387
ချန်ပီယံ	0.50700
ကလပ်	0.49973
ဆီမီးဖိုင်နယ်	0.46233

Table 5. (b) Health Terms

Health Category	
Term	**TF-IDF**
ကင်ဆာ	0.49203
သုတေသီ	0.48936
ရာခိုင်နှုန်း	0.46564
သုတေသန	0.44441
နှလုံး	0.4405
ကိုယ်ဝန်ဆောင်	0.4189
လက္ခဏာ	0.39852
မှတ်ဉာဏ်	0.38696

Table 5. (c) Crime Terms

Crime Category	
Term	**TF-IDF**
အမှု	0.66555
ပုဒ်မ	0.59743
ယူဆောင်	0.53251
ထွက်ပေး	0.5286
ခရီးသည်	0.47805
မီးသတ်	0.47805
အခင်းဖစ်	0.47506
ပြက္ခဒိန်	0.47200

In this study, to evaluate the performance of feature selection process, top 100 features are selected as positive tuples (important features) and 300 features that have lowest TF-IDF scores are selected as negative tuples (not important features to be removed). To check the accuracy of selected features, domain specific lexicons that contains the most widely used words about 350 words for each topic are manually constructed. Table 6. and 7. show the evaluation measure by TF-IDF (Baseline Method) and TF-IDF (Adaptive Method). As TF-IDF scores of the features that only contain in one domain increase and TF-IDF scores of the features that have much term frequency but contain in many domains (that can cause multi-label problem) decrease in TF-IDF (Adaptive Method), the performance of adaptive method is better than baseline method. According to the experiment, adapted TF-IDF model shows an improvement in performance than existing method.

Table 6. Evaluation Measure by TF-IDF (Baseline Method)

	Precision (%)	Recall (%)	F-Measure (%)	Error Rate(%)	Accuracy (%)
Crime	91	51	65	14	87
Sport	97	68	79	9	91
Health	91	73	83	11	91

Table 7. Evaluation Measure by TF-IDF (Adaptive Method)

	Precision (%)	Recall (%)	F-Measure (%)	Error Rate(%)	Accuracy (%)
Crime	96	59	73	11	89
Sport	100	77	87	5	94
Health	97	82	83	8	95

Table 8. describes the formulas of each measure and terms used in these formulas are briefly described [1].

Table 8. Evaluation Measure

Measure	Formula	Description
Accuracy	$\frac{TP + TN}{P + N}$	**TP**- True Positives refer to positive tuples correctly labeled
Precision	$\frac{TP}{TP + FP}$	**TN**- True Negatives refer to negative tuples correctly labeled
Recall	$\frac{TP}{TP + FN} = \frac{TP}{P}$	**FP**-False Positive refer to negatives tuples that were incorrectly labeled as positive
F-Measure	$\frac{2 * Precision * Rec}{Precision + Reca}$	**FN**-False Negative refer to positive tuples that were mislabeled as negative
Error Rate	$\frac{FP + FN}{P + N}$	***P*** **–**the number of positive tuples ***N*** -the number of negative tuples

9. Error Analysis

The problem with TF-IDF method is that the ranges of TF-IDF scores for each domain are not on the same scale. The domains with large number of extracted features have higher TFIDF values than the other domains with the smaller number of features. So, it is needed to justify the number of input features for each domain in order to get reasonable scores. Then, the problem of out of dictionary words, for instance, the word "ပယ်နယ်" has high tf-idf score in sport domain because the word "ပယ်နယ်တီ" is often used in sport news and the word "ပယ်နယ်တီ" does not contain dictionary. In this case, such kinds of words are added to Myanmar words list but some of the words cannot be noticed.

10. Conclusion and Future Works

This study especially focuses on feature selection and the aim of this study is to select high informative features from collection of online news documents for future online news categorization process. Myanmar news features are selected by TF-IDF (Baseline Method) and TF-IDF (Adaptive Method). The experimental results show the higher performance of adaptive method than baseline method. Further experimental work will be performed with cosine similarity on latent semantic analysis (LSA) vectors, the Latent Dirichlet Allocation (LDA) model and other feature selections methods then more categories of online news articles will be considered in the future.

9. References

[1] Han, Jiawei and Kamber, Micheline and , and Pei, Jian "Data Mining: Concepts and Techniques (Third Edition)", Morgan Kaufmann, Third Edition, The Morgan Kaufmann Series in Data Management Systems, Boston, 2012, pp. 365-366.

[2] Tun Thura Thet, Jin-Cheon Na, Wunna Ko Ko, "Word Segmentation for the Myanmar language", J*ournal of Information Science 34 (5)*, 2008, pp. 688-704.

[3] Dadgar, Seyyed Mohammad Hossein, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. "A Novel text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification." *In Engineering and Technology (ICETECH), 2016 IEEE International Conference on*, IEEE, 2016, pp. 112-116.

[4] Islam, M.S., Jubayer, F.E.M. and Ahmed, S.I., "A Support Vector Machine Mixed with TF-IDF Algorithm to Categorize Bengali Document". *In Electrical, Computer and Communication Engineering (ECCE), International Conference on*, IEEE, February 2017, pp. 191-196.

[5] Liu, M. and Yang, J., "An Improvement of TF-IDF Weighting in Text Categorization". *International Proceedings of Computer Science and Information Technology*, 2012, pp.44-47.

[6] Tin Htay Hlaing, "Manually Constructed Context-Free Grammar for Myanmar Syllable Structure". *In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 32-37.

[7] Z. M. Maung, Mikami Yoshiki, "Rule-based Syllable Segmentation of Myanmar Texts*". In Proceedings of the 6 thWorkshop on Asian Language Resources*, January 2008, Hyderabad, India, pp. 11-12.

[8] K.Stribley, "Collation of Myanmar in Unicode", *technical report*, June 17, 2007.

[9] https://boilerpipe-web.appspot.com

[10] https://en.wikipedia.org/wiki/Tf-idf

[11] https://github.com/ye-kyaw-thu/sylbreak

[12] http://news-eleven.com/

[13]https://raw.githubusercontent.com/lwinmoe/segment/master/burmese-word-list.txt

[14]https://thanlwinsoft.github.io/www.thanlwinsoft.org/ThanwinSoft/MyanmarUnicode/Conversion/myanmarConverter.html

[15] http://thithtoolwin.mmbloggers.com