Review Paper:

A Review on Text Categorization

Nguyen Thi Xuan Trang
University of Economics, The University of Danang, Danang, VIETNAM
trangntx@due.edu.vn

Abstract

Text categorization is one of the most important text mining tasks to deduce patterns for certain targets from a document-based intermediate form. This study reviews almost basic contents of text categorization: major stages, classification techniques and empirical researches on applying this tool for documents written in various kinds of languages.

After reviewing previous researches on this topic, it was found that the effectiveness of Support Vector Machines method was confirmed in most empirical researches on text categorization for different languages such as English, Arabic, Korean or Vietnamese. The study also suggests for future researches on exploring an effective process of text categorization that can be applied for documents in any kind of language.

Keywords: Text mining, text categorization, language

Introduction

Due to huge amount of information available on the Internet with an unstructured text format, text mining emerged as a practical application of several mechanisms extracting interesting and useful knowledge from such chaotic text collections; thus, making it easier for users to find the information they need. One of the most important text mining tasks is text categorization.

This study will summarize basic contents of text categorization such as its definition, its main stages, techniques or classification algorithms. Additionally, it will also review empirical researches on text categorization for different languages to comprehend how text categorization was implemented in practice.

Review of Literature

An overview on text categorization: According to Kantardzic[2], text categorization or text classification was one of text-analysis tools in a text-mining framework (Figure 1). Kantardzic[2] defined text mining as a process by which unstructured text data would be analyzed to extract useful information for different purposes. This process was suggested to include two phases: text refining that transformed text, called an entity, from a free form into a document-based intermediate form or concept-based intermediate form depending on whether that entity represented a document or an object respectively and *knowledge distillation* that inferred semantic patterns and relationships from the chosen intermediate form.

Figure 1 listed some operations for mining a concept-based intermediate form such as predictive modeling and associative discovery, or mining a document-based intermediate form, for example, clustering, categorization and visualization. Therefore, text categorization was one of tools to deduce patterns for particular targets from a document-based intermediate form.

In the literature review of text categorization, most researchers were interested in automatic text categorization or text classification. Automatic text classification can be defined as a process where an unstructured text document was automatically classified according to its desired categories.[1]

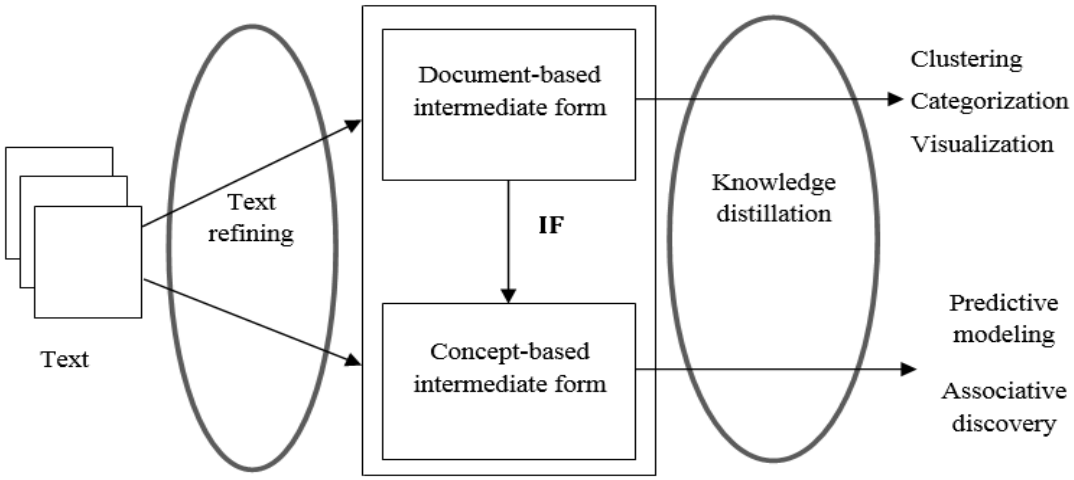

Figure 1: A text-mining framework[2]

Nejad et al[5] suggested three main stages of text categorization: preprocessing, feature selection and classification. They mentioned that it would be necessary to preprocess the text document and then select significant features of the text before classifying the preprocessed text by different learning algorithms such as Naïve Bayes, Decision Tree or Support Vector Machine.

After reviewing preceding researches, Odeh et al[7] summarized five text categorization steps including document preprocessing, document representation, dimension reduction, model training and testing and evaluation as the illustration of figure 2. In comparison with the suggestion of Nejad et al[5], the feature selection stage was divided into two steps: document representation and dimension reduction; in the meanwhile, the classification stage was also proposed to separate into two steps: model training and testing and evaluation by Odeh et al.[7]

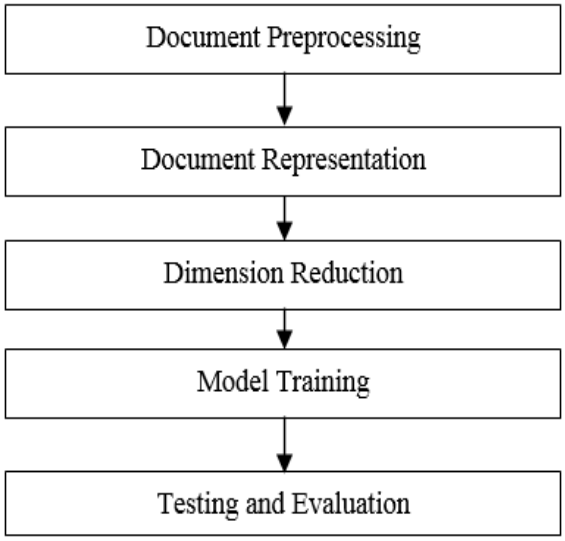

Figure 2: Text categorization steps[7]

As per the opinion of Odeh et al[7], the step document preprocessing first removed html tags, rare words as well as stop words and applied stemming in case this process was necessary. After that, this step normalized remaining words in the document owing to different root extraction methods. Next, the purpose of the step *document representation* was to transform the document into an identifiable format. For instance, a document was considered as a multi-dimension vector and each feature of the document was regarded as a dimension of the vector in vector space model. The third step was dimension reduction to opt for the most meaningful and representative features of the document.

The following step, model training, was evaluated to be the most important step consisting of selecting a training set, learning on that training set and then producing an appropriate model. Based on this model, classification would be performed and proper indices would also be selected for evaluations of the classification in the last step, testing and evaluation.

The extent of economies of scope: Mahalakshmi and Duraiswamy[4] listed seven basic classification techniques for text documents: Back propagation Neural Network, Latent semantic indexing, Support vector machines, Decision trees, Naive Bayes classifier, Self-Organizing Map and Genetic Algorithm. First, the Back propagation Neural Network technique is applied to train neural networks with several layers from an input layer to an output layer.

Depending on the input pattern that was given to the input layer, the network will extract the output that corresponds with the desired output pattern. Secondly, the Latent semantic indexing technique based on mathematics established relationships between the terms and concepts and then to identified patterns.

In terms of Support vector machines, it is suggested to be a dominant learning algorithm utilized for text categorization. For prediction purpose, a model will be built from a given set of training examples thanks to Support vector machines training algorithm. Similarly, a decision tree, one of inductive learning algorithms, can be seen as a predictive model that expresses observations for an item. Its leaves and branches represent classifications and conjunctions of features respectively. Next is Naive Bayes classifier technique. With the assumption that there was the independence among features of a class, the Naive Bayes classifier applies Bayes theorem to calculate essential parameters for classification.

Finally, while Self-Organizing Map includes a set of models that illustrate input data through a similarity graph, Genetic Algorithm is a useful technique for global searching with three principles: "biological evolution, natural selection and genetic recombination".[4]

In addition to various approaches to text categorization that Mahalakshmi and Duraiswamy[4] mentioned, Pawar and Gawande[8] added some other techniques such as K. Nearest Neighbor, Rochio's Algorithm, Neural Text Categorizer Acronyms and Soft-Supervised Learning. Moreover, they introduced hybrid approaches which combined several classifiers for text categorization; for example Neural Networks Initialized with Decision Trees approach that was the combination between the Back propagation Network technique and the Decision Trees technique, or Probabilistic Neural Network approach that combined three different techniques (Support vector machines, K. Nearest Neighbor and Decision Trees), or *Bahes* Formula for Classification that linked between Naïve Bayes Algorithm and Support Vector Machines method.

Pawar and Gawande[8] also compared the effectiveness of various classifiers for Reuters 21578 and 20 Newsgroup Datasets that were reported by previous researchers and they recognized that Support Vector Machines was one of the most effective text classification method compared with other supervised classification algorithms.

Empirical researches on text categorization for different languages

Arabic text categorization: Increasing necessity of automatically organizing unstructured text documents in its desired category appears to be more challenging when it comes to Arabic contexts. The research of Hmeidi et al[1] was concerned with TC of Arabic articles, comparing the five best known learning algorithms (classifiers): Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (J48) and Decision Table.

Hmeidi et al[1] also studied the effect of stemming on Arabic TC by using two versions of the dataset which have been pre-processed through light and root-based stemmers in addition to the original version on which no stemming technique was applied. They conducted the research with a sample of 1,000 documents equally spread across 5 categories (Economics, Health, Politics, Sports and Technology), using the 5-fold cross-validation for testing with 4 popular measurements: accuracy, precision, recall and F1-score. The results illustrated that stemming in general could considerably reduce the learning times (model building) of many classifiers that would have had unacceptably long learning times otherwise.

Furthermore, SVM was shown to achieve the highest accuracy, followed closely by NB, regardless of the stemming technique whereas the light10 stemmer generated better general results than the root-based stemmers and no stemming at all. A comparison between two popular TC tools Weka and RapidMiner suite was also performed to discover whether different results and observations could be obtained from these tools using the same dataset based on the evaluation of accuracy and scalability. Since the Weka tool was troublesome for some reasons because of limited memory capability, Arabic TC researchers were recommended to utilize RapidMiner owning to its effectiveness and scalability.[1]

Odeh et al[7] also researched on text categorization for Arabic documents. They suggested vector evaluation method for this process with seven steps (Figure 3). This method was started with removing Arabic stops words from the documents before normalizing the remaining words. The next step was applying stemming to transfer inflected words into their root forms. After that, two words with the highest weights would be selected after calculating the weights of entirely words in the tested document based on weighting scheme function with the appearance of the term occurrence frequency and the inverse document frequency. These two words would be compared with key words of the corpus categories to realize the most appropriate category that would be returned its name in the last step. Odeh et al[7] supported the vector evaluation method because they found that its categorization could provide high precision rate.

English text categorization: Nejad et al[5] proposed the process of text categorization as figure 4 with three main stages: preprocessing that contained five steps (Transform Case, Tokenization, Filter Stop Words, Stemming and Generate n-Gram), feature selection and classification through LIBSVM that was an integrated software supporting multi-class classification owing to various Support Vector Machines formulations. Their suggested method was implemented on reuters-21578 dataset with 7674 text documents comprising 5485 train documents and 2189 test documents. Nejad et al[5] compared this method with Naïve Bayes and J48 according to five evaluation criteria: Accuracy, Classification Error, Precision, Recall and F1. Finally, they confirmed the superiority of their proposed method with the highest accuracy, precision, recall as well as F1 and the lowest classification error among three methods they applied.

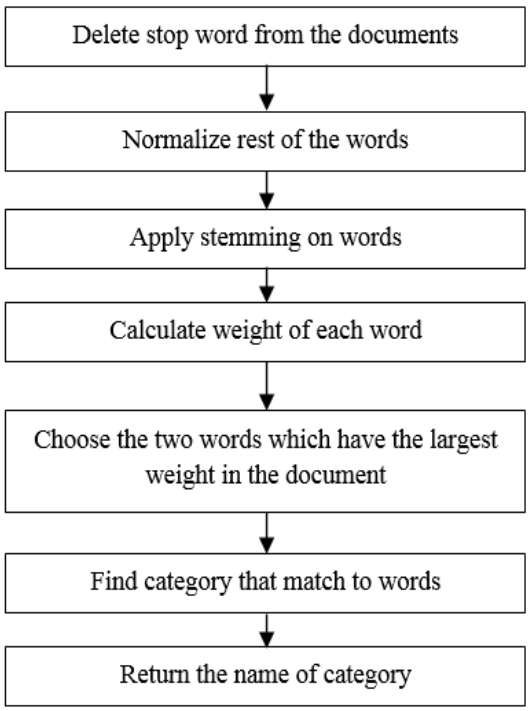

Figure 3: Arabic text categorization steps[7]

Korean text categorization: In 2004, Ko et al[3] recommended a new indexing method for text categorization illustrated in figure 5. They suggested that after preprocessing stage, the importance of sentence should be calculated by using the title or the importance of terms before performing indexing. Subsequently, four types of Text Classifiers (Naïve Bayes, kNN, Rocchio, SVM) were utilized to assign proper categories. Ko et al[3] collected 10,331 Korean documents containing 7,224 documents for training data and 3,107 ones for test data in order to test their proposed system.

The results showed that the classification performance of their text categorization system became better than that of the basis system in all four classifiers. Hence, they emphasized on the usefulness of two text summarization techniques (use the title or use the importance of terms) to measure the importance of sentences in the process of text

categorization. It was noticeable that their conclusions were also confirmed on English documents – the second newsgroup data set in the research.

Nguyen Linh Giang and Nguyen Manh Hien[6] tested the effectiveness of Support Vector Machine (SVM) method in categorizing documents in Vietnamese. The used sample was 4162 documents among which 50 percent of the documents belonged to training data and 50 percent left was for test data. After applying SVM for this sample, they found that the accuracy of this classification technique was acceptable at 80.72% in practice in Vietnam.

In order to increase its accuracy, it was suggested that succeeding researchers should find out either how to improve the preprocessing stage or how to adjust algorithms for training SVM.

Conclusion

This study can be considered as a basic reference for beginning learners who are interested in text categorization in text mining framework. It reviewed almost basic contents of text categorization: major stages, classification techniques and how it was applied for categorizing documents written in different languages in practice. It was recognized that there has been two main research streams in this topic with the final purpose of improving the classification performance.

The first stream was trying to discover the most effective text classification techniques by comparing the performance of different text classifiers through various measurements such as accuracy, precision, recall, classification error and F1-score. Most empirical studies supported the effectiveness of Support Vector Machines method compared with other classification algorithms.

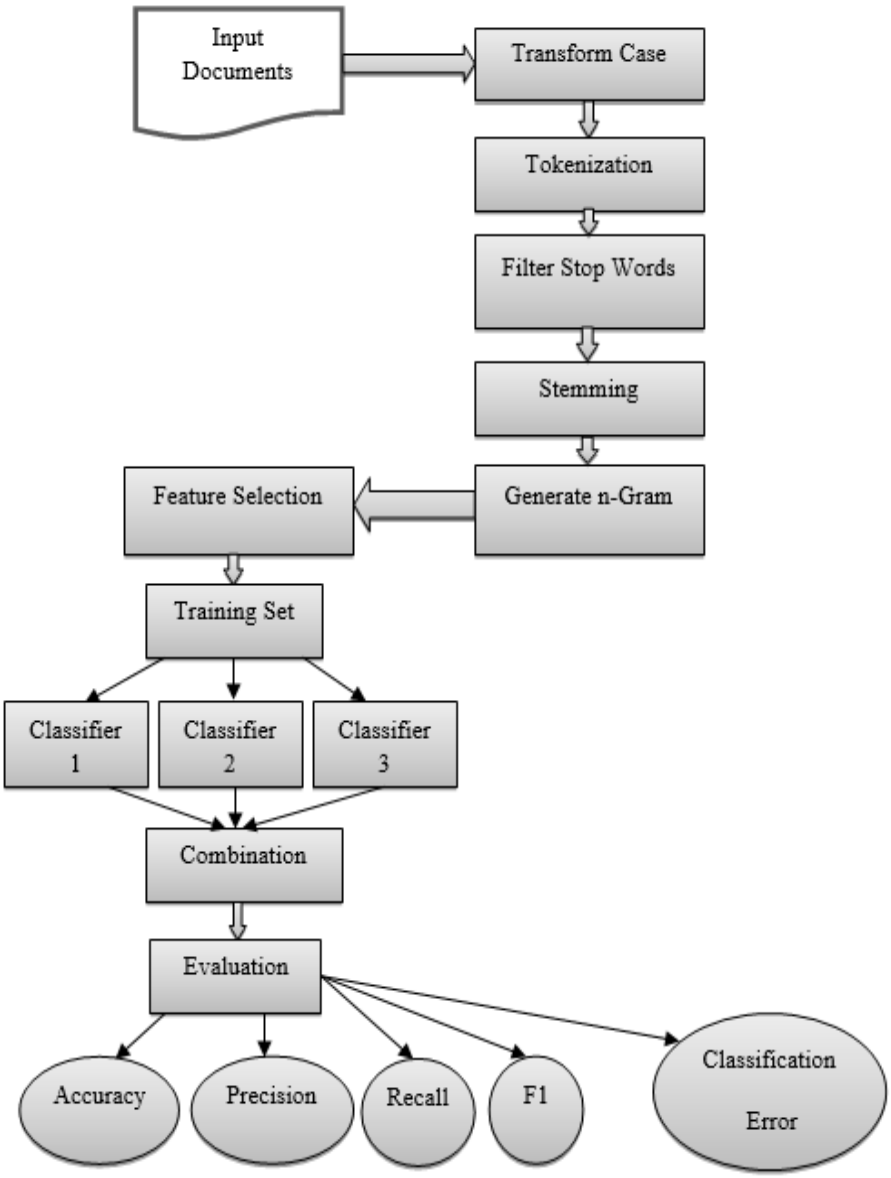

Figure 4: Proposed text categorization system as opinion of Nejad et al[5]

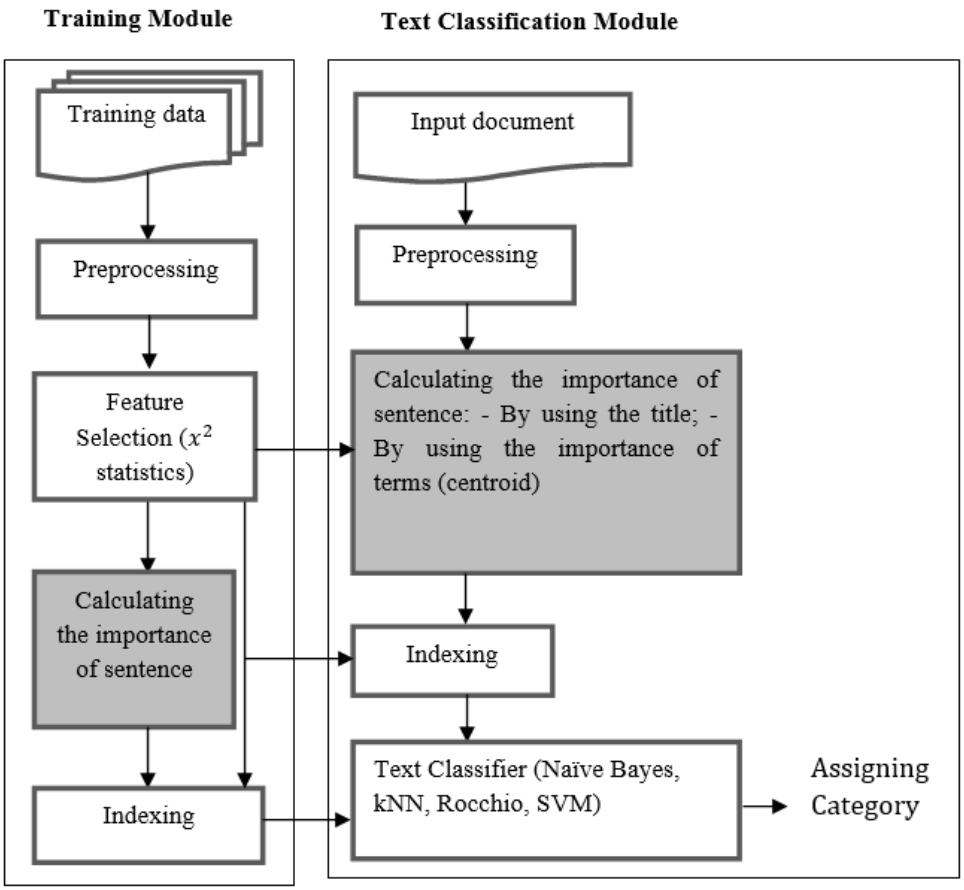

Figure 5: Proposed text categorization system as opinion of Ko et al[3]

The second research stream was adjusting detailed steps in each stage of text categorization: preprocessing, feature selection, or classification to find out an effective process of text categorization. For instance, Ko et al[3] proposed to apply two text summarization techniques (use the title or use the importance of terms) to measure the importance of sentences in the feature selection stage of text categorization. To this research stream, each document written in each type of language such as English, Arabic, Korean or Vietnamese may possess its own effective process of text categorization because of own special features of each language. Therefore, it is suggested that future researches can compare the application of their proposed processes in documents written in different languages.

References

1. Hmeidi I., Al-Ayyoub M., Abdulla N.A., Almodawar A.A., Abooraig R. and Mahyoub N.A., Automatic Arabic text categorization: A comprehensive comparative study, *Journal of Information Science*, **41(1)**, 114–124 **(2015)**

2. Kantardzic M., Data Mining: Concepts, Models, Methods and Algorithms, 2nd edition, John Wiley and Sons, Inc.: Institute of Electrical and Electronics Engineers **(2011)**

3. Ko Y., Park J. and Seo J., Improving text categorization using the importance of sentences, *Information Processing and Management*, **40**, 65-79 **(2004)**

4. Mahalakshmi B. and Duraiswamy Dr. K., An Overview of Categorization techniques, *International Journal of Modern Engineering Research*, **2(5)**, 3131-3137 **(2012)**

5. Nejad M.B., Attarzadeh I. and Hosseinzadeh M., An Efficient Method for Automatic Text Categorization, *International Journal of Mechatronics, Electrical and Computer Technology*, **3(9)**, 314-329 **(2013)**

6. Nguyen Linh Giang and Nguyen Manh Hien, Classification of Vietnamese Documents Using Support Vector Machine (Phân loại văn bản tiếng Việt với bộ phân loại vector hỗ trợ SVM), *Journal of Postal Telecommunications and Information Technology*, **15**, 66-75 **(2005)**

7. Odeh A., Abu-Errub A., Shambour Q. and Turab N., Arabic text categorization algorithm using vector evaluation method, *International Journal of Computer Science and Information Technology*, **6(6)**, 83-92 **(2014)**

8. Pawar P.Y. and Gawande S.H., A Comparative Study on Different Types of Approaches to Text Categorization, *International Journal of Machine Learning and Computing*, **2(4)**, 423-426 **(2012)**.

(Received 13th June 2022, accepted 05th August 2022)