Text Categorization with Text Summarization Techniques

P.Vijaya Pal Reddy

Matrusri Engineering College, Hyderabad

Abstract

This paper addresses the issue of curse of dimensionality in the Text Classification (TC) problem using Text Summarization (TS). In this paper, an attempt is made to effectively tackle the curse of dimensionality problem using Text classification model. It is observed from the experimental results that proposed model improves the performance of text categorization. An empirical evaluation is performed on different summaries and it is concluded that performance of the TC is observed to be improved when the redundancy in the summary is minimal. Performance of the proposed model is evaluated with the F1measure and macro-averaged F1 measures using Support Vector Machine (SVM).

Keywords: *Text classification, Text summarization, Support Vector Machines, Term Frequency, sentence location, Centrality.*

1. Introduction

Now a day's huge amount of information is being posted on to the web. In order to get useful information from the web, the information available has to be categorized. The task of classifying given data into a prespecified set of categories is known as "Text Categorization" (TC). In the process of TC, feature selection methods are used to reduce the dimensionality known as curse of dimensionality of the dataset. Text summarization is a technique which extracts the important information from a source(s) to produce an abridged version for a particular user or task [8]. Summary of a document can be generated by assigning a score to each sentence. From the literature survey, it is observed that some authors [1, 2, and 3] has performed TC by applying summarization techniques. In this paper, we addressed the problem of curse of dimensionality using TS approach to minimize noisy word count in the feature set presented for TC.

In this paper, Section 2 briefs about different studies on text classification, text summarization and TC using TS. Section 3 explains about the different phases in the proposed model and also about the method of extractive summary generation. Description about dataset and analysis on results achieved are presented in section 4. The final section gives the conclusions from the results.

2. Related Work

Text Categorization is a process of assigning unlabeled document to a predefined category(s). TC techniques can be broadly categorized into supervised, unsupervised and semi supervised approaches. Naive Bayes [4, 5], Rocchio [6], K-Nearest Neighbor [7], decision trees [8, 9], support vector machines (SVM) [10] are the most prominent algorithms in supervised learning process. From these approaches SVM outperforms compared with other methods [11,12,13]. Hence in this paper, we used SVM as the classier for TC. Dimensional space of a Telugu document is large when compared with English text. Performance of a classifier depends on the feature set which is deduced from the vector space of the document. Classifiers performance are improved with an optimal feature set from the vector space proved in [15, 16, 17].

There are many methods proposed for feature set selection [24] of a dataset. Though there are many methods proposed for feature set selection, there is fair chance of including noisy features in the feature set. By identifying topic sentences i.e summary of the document, noisy features can in the feature set can be greatly reduced. Feature set generation from the summary of the document instead from the original document not only reduces the noisy features, but also saves the time of the TC process.

Text Summarization (TS) is a technique which extracts the important information from a text document(s) for the user [14]. In TS, sentences are ranked according to their relevance to the document and extracts the sentences which are more relevant for the document to form a summary until the all the topics in the document are covered without redundancy [18,19]. The score or the relevance of the sentence can be calculated based on different features like syntactic, semantic and combination of these features [20, 21]. There are many researchers worked on Test classification based on Text summarization. Most relevant features are selected by using summarization and a classifier is build based on the reduced feature space [2]. In Ko [1], summarization technique is applied on all the features and are re-weighted based on the importance of sentences. Performance of web page classification is employed by Shen [3] using summarization technique. In this paper an attempt is made to show that the Text Classification performance is improved by using summary of the document on Telugu text. In this paper, we considered research on Hard Categorization.

3. Proposed Model

All The proposed model contains four phases as shown in the Figure 1 such as preprocessing, text summarization, feature weighting and classification. The functioning of these phases are briefly described below.

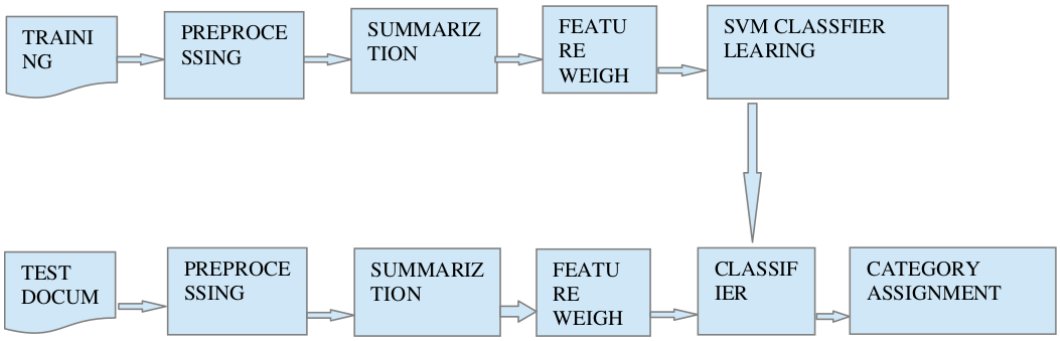

Figure 1: Text Categorization Model Using Text SummarizationA. Preprocessing

Document need to be preprocessed before processing through the machine. The pre processing phase contains removal of unnecessary content from the document which is not so useful for TC which look like punctuation marks, numbers, dates and symbols etc. Secondly, features which can create noise to the TC process called stop words which are used to give meaning to the sentence and it is necessary to remove these stop words. For Telugu text stop words are identified by using the NLTK tool kit. As Telugu is complex morphological variant language, reducing the features of document into their root form can greatly reduce the dimensionality space of the document. Hence features of the document are converted into their root form using Telugu Morphological Analyzer (TMA) tool. After preprocessing the feature space of the document contains only stemmed form of the features.

B. Text Summarization

Summarization can be broadly divided into two categories namely extractive summarization and abstractive summarization. An extractive summarization is a process of selecting a set of sentences from the original document which gives the gist of the document, while an abstractive summarization is a reformulation of the original document [22] probably with new sentences. In this paper, an extractive summarization is used because of simplicity, robustness and domain independence. A combination of three surface features for summary generation such as Term frequency of a sentence in the document (TF), Sentence Location in the document (SL) and Centrality of the sentence in the document (CE) were used. The formula for calculating these features and finding the suitable combination of these features are presented below:

i. Term Frequency (TF)

The term frequency of a word in a document is the number of occurrences of the word in that document. This count is normalized to overcome a bias towards longer documents. Term frequency of a word is calculated as follows:

$$tf_i = \frac{c_i}{\sum_k c_k}$$

where 'Ci ' is the number of occurrences of a word in the document, and the denominator is the total number of words in the document. The term frequency score of a sentence 'S' is calculated as:

$$Score_{f_1}(S) = \frac{\sum_{k=1}^{n} tf_k}{|S|}$$

Where numerator gives the sum of all the word frequencies in the sentence S and denominator represents the number of words in the sentence.

ii. Sentence Location (SL)

Sentences at the beginning of the texts of news documents gives the general information of the document which are suitable to form a gist. The remaining sentences of document are the details about the news which has less importance to include in the gist. Therefore important sentences, for the gist, are usually located at some particular positions. In order to formalize the sentence location, each sentence is given a location value Li (Li is equal to i). Then in order to give higher score to the first sentences, we use the formula mentioned below which gives the position score of a sentence S. where 'R' is the number of sentences in the document

iii. Centrality (CE)

The centrality of a sentence implies its similarity to other sentences, which is measured as the degree of overlapping between sentences to other sentences. If a sentence has high centrality means that the sentence introduce many topics of the document. Therefore, high centrality sentences are more preferable in summary than low centrality sentences. The formula to find the score of centrality of a sentence 'S' is:

$$Score_{f_3}(S) = \frac{|wordsofS \cap wordsofremainingsentences|}{|wordsofS \cup wordsofremainingsentences|}$$

iv. Summary Generation For a sentence 'S', the weighted score function combine all the feature scores of the sentence as follows.

$$Score(S) = \sum W_i \times Score f_i(S)$$

Where W i represents the weight assigned to feature 'i' to generate the summary that best expresses the gist of the document. All possible weight combinations between 0 and 1 with an interval of 0.1 is evaluated of the training dataset. From the empirical evaluations it is concluded that the best weights for TF, CE and SL are 0.2, 0.3 and 0.5 respectively to generate more appropriate gist of the document. Then the sentences are ranked according to the above weights and are assigned to each feature.

C. Feature Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) approach for feature weighing is carried out which is proved to be more prominent in the TC literature. The product of TF and IDF is given as:

$$W(d,t) = TF(d,t).IDF$$

where TF(d,t) is the number of times the term t occurs in the document d

$$IDF(t) = log$$

where N is the total number of documents in the training dataset and n is the number of documents that contain the term t.

4. Experimental Analysis

A. Document Collection

The dataset was gathered from Telugu News Papers such as Eenadu, Andhra Prabha and Sakshi from the web during the year 2009 – 2010. The corpus is collected from the website http://uni.medhas.org/ in unicode format. We obtained around 800 news articles from the domains of Economics, Politics, Science, Sports,Culture and health. Before proceeding, preprocessing like tokenisation is carried out by removing stop words . We have chosen 60% of the documents as training dataset, remaining 40% of the documents as testing dataset for all six categories. Using combined sentence score, we have generated the scores for all sentences for for both training dataset and testing dataset. The experiments were carried out with TF-IDF weighing method using the SVM classifier for various levels of compressions on the original document. We have compared the performance of the SVM classifier among various levels of compressions of the documents with original document using F1 measure and macro-averaged F1 measure.

B. Evaluation Measures

To evaluate the classification performance of the proposed model using SVM classifier, we used F1 measure and macro-averaged F1 measure. F1 measure and macro-averaged F1 measures are calculated as follows:

$$F_1 = \frac{2 * Recall * Preci}{Recall + Precis}$$

$$Precision = \text{ and}$$

$$Recall = \frac{X}{X + Z}$$

where X is documents retrieved and relevant, Y is documents retrieved but not relevant and Z is documents not retrieved but relevant. Macro-averaged F-measure is obtained by taking the average of F-measure values for each category as:

$$F(macro-average) = \frac{\sum_1^M F_i}{M}$$

We have compared the performance of the proposed model with classification performance on original dataset using SVM. For the original classification performance, we have used CHI-SQUARE as the feature selection measure and TF-IDF as the feature weighting method. Feature set is formed with first 100 best features from the vector space of the documents.

Table 1: Comparison among different levels of summaries with original Text classification model

Category	Original	10%	20%	30%	40%	50%
Economics	0.733	0.751	0.772	0.786	0.763	0.741
Politics	0.828	0.838	0.858	0.882	0.864	0.850
Science	0.751	0.747	0.774	0.805	0.793	0.778
Sports	0.890	0.912	0.935	0.944	0.920	0.893
Culture	0.843	0.839	0.862	0.896	0.857	0.849
Health	0.909	0.923	0.940	0.967	0.938	0.927
Macro-averaged F1	0.826	0.835	0.867	0.880	0.856	0.840

From the results we observed that classification performance was improved through the proposed model compared with traditional Text categorization model. The probable reason is summary can retain the important information of the original document and also reduces the number of features which can create noise in the process of classification learning by the classifier. The 30% summary of the original document gives the best performance with 5.6% improvement in the macro-averaged F1performance when compared with the original macro-averaged F1 value. For more compression levels i.e upto 20% of the summary, there a probability of missing some content features which are more useful for categorization. This causes less performance in classification. Similarly for less compression levels i.e from 40% and above of summary, it includes some of the features which acts as noisy words in the classification process leads to decrease in the classification performance.

5. Conclusions

A document contains set of sentences. A few sentences explains about the topics covered in the document. The remaining sentences are act as supporting sentences to support the explanation of the topics. Summary of a document contains the sentences are used to explain about the topics covered in the document which can be called as topic sentences. Hence the summary of the document retains important features of original document and also reduces the noisy features from the document which increases the performance of the classifier.

References

[1] Y. Ko, et al., "Automatic text categorization using the importance of sentences," in Proceedings of the 19th International Conference on Computational Linguistics, Vol. 1, 2002, pp. 1-7.

[2] A. Kolcz, et al., "Summarization as feature selection for text categorization," in Proceedings of the 10th International Conference on Information and Knowledge Management, 2001, pp. 365-370.

[3] D. Shen, et al., "Web-page classification through summarization," in Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, 2004, pp. 242-249.

[4] A. McCallum and K. Nigam, "A comparison of event models for Naïve Bayes text classification," in Proceedings of AAAI Workshop on Learning for Text Categorization, 1998, pp. 41-48.

[5] R. R. Yager, "An extension of the naïve Bayesian classifier," Information Sciences, Vol. 176, 2006, pp. 577-588.

[6] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 143-151.

[7] I. Rahal and W. Perrizo, "An optimized approach for KNN text categorization using P-trees," in Proceedings of ACM Symposium on Applied Computing, 2004, pp. 613-617.

[8] E. Gabrilovich and S. Markovitch "Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5," in Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 321- 328.

[9] T. Hidekazu, et al., "Estimating sentence types in computer related new product bulletins using a decision tree," Information Sciences, Vol. 168, 2004, pp. 185-200.

[10] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[11] S. Dumais, et al., "Inductive learning algorithms and representations for text categorization," in Proceedings of the 7th International Conference on Information and Knowledge Management, 1998, pp. 148-155.

[12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proceedings of the 10th European Conference on Ma chine Learning, 1998, pp. 137-142.

[13] Y. Yang and X. Liu, "A re-examination of text categorization methods," in Proceedings of Special Interest Group on Information Retrieval, 1999, pp. 42-49.

[14] Y. Guo and G. Stylios, "An intelligent summarization system based on cognitive psychology," Information Sciences, Vol. 174, 2005, pp. 1-36.

[15] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and Naïve bayes," in Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 258-267.

[16] M. Rogati and Y. Yang, "High-performing feature selection for text classification," in Proceedings of the 11th International Conference on Information and Knowledge Management, 2002, pp. 659-661.

[17] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 412-420.

[18] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 10-17.

[19] D. Marcu, "Discourse trees are good indicators of importance in text," Advances in Automatic Text Summarization, 1999, pp. 123-136.

[20] H. P. Edmundson, "New methods in automatic extracting," Journal of the Association for Computing Machinery Vol. 16, 1969, pp. 264-285.

[21] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research and Development, Vol. 2, 1958, pp. 159-165.

[22] E. Hovy and C. Y. Lin, "Automated text summarization in SUMMARIST," Advances in Automatic Text Summarization, 1999, pp. 81-94.

[23] T. Joachims, "Making large-scale support vector machine learning practical," Advances in Kernel Methods: Support Vector Machines, 1999, pp. 169-184..

[24] YANG Y, PEDERSEN J Q. A comparative study on feature selection in text categorization. Proceedings of the 14th