International Journal of Scientific Research in Computer Science and Engineering
Research Paper Volume-3, Issue-2 ISSN: 2320-7639

Evaluation of Stemming and Stop Word Techniques on Text Classification Problem

Dharmendra Sharma[1]* and Suresh Jain[2]

[1*,2]Mewar University, Chittorgarh, Rajasthan, India

Available online at www.isroset.org

Abstract— Now-a-days a huge amount of information is available over the internet in electronic format. This large amount of data can be analyzed to maximize the benefits, for intelligent decision making. Text categorization is an important and extensively studied problem in machine learning. The basic phases in text categorization include preprocessing features, extracting relevant features against the features in a database, and finally categorizing a set of documents into predefined categories. Most of the researches in text categorization are focusing more on the development of algorithms for optimization of preprocessing technique for text categorization. In this paper we are summarizing the impact of stop word and stemming onto feature selection.

Keywords—machine learning, stemming, feature selection

I. INTRODUCTION

Amazing development of Internet and digital library has triggered a lot of research areas. Text categorization is one of them. Text categorization is a process that group text documents into one or more predefined categories based on their contents [1]. It has wide applications, such as email filtering, category classification for search engines and digital libraries. Associative text classification, a task that combines the capabilities of association rule mining and classification, is performed in a series of sequential subtasks. They are the preprocessing, the association rule generation, the pruning and the actual classification. Out of these, the first step, that is, 'Preprocessing', is the most important subtask of text classification. The importance of preprocessing is emphasized by the fact that the quantity of training data grows exponentially with the dimension of the input space. It has already been proven that the time spent on preprocessing can take from 50% up to 80% of the entire classification process [2], which clearly proves the importance of preprocessing in text classification process. This paper discusses the various preprocessing techniques used in the present research work and analyzes the affect of preprocessing on text classification using machine learning algorithms. Section 2 gives an overview of the work in text preprocessing. Section 3 explains the preprocessing steps used. Experimental results are described in section 4. Summarization of work narrated in Section 5.

II. RELATED WORK

The preprocessing phase of the study converts the original textual data in a data-mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified. It is the process of incorporating a new document into an information retrieval system. An effective preprocessor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance (precision and recall). This phase is the most critical and complex process that leads to the representation of each document by a select set of index terms. The main objective of preprocessing is to obtain the key features or key terms from online news text documents and to enhance the relevancy between word and document and the relevancy between word and category. Our method is the evaluation of the weighting methods. Until now, there are many researches about weighting method. The reference [3] describes survey about the weighting methods such as binary [4], term frequency (TF) [4], augmented normalized term frequency [4] [5], log [5], inverse document frequency (IDF) [5], probabilistic inverse [4] [5], document length normalization [4].

III. METHODOLOGY

The goal behind preprocessing is to represent each document as a feature vector, that is, to separate the text into individual words. In the proposed classifiers, the text documents are modeled as transactions. Choosing the keyword that is the feature selection process, is the main preprocessing step necessary for the indexing of documents. This step is crucial in determining the quality of the next stage, that is, the classification stage. It is important to select the significant keywords that carry the meaning, and discard the words that do not contribute to distinguishing between the documents. The procedure used for preprocessing the web document dataset is shown in Fig.1.

Step 1:Data Collection
Step 2: Stop word removal
Step 3: Stemming
Step 4: Indexing
Step 5: Term weighting
Step 6: Feature Selection

Fig. 1: Processing steps

ISROSET- Int. J. Sci. Res. in Computer Science & Engineering Vol-3(2), PP (1-4) Apr 2015, E-ISSN: 2320-7639

A. Data Collection

For the data set we use Google web API to collect the document. We have collected 64 documents on ipad concept it has 9998 features.

B. Stop Word Removal

Stop-words are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). In English language, there are about 400- 500 Stop words. Examples of such words include 'the', 'of', 'and', 'to'. The first step during preprocessing is to remove these Stop words, which has proven as very important [6]. Many of the most frequently used words in English sentence are useless in Information Retrieval (IR) and text mining.

C. Stemming

Stemming techniques are used to find out the root/stem of a word. Stemming converts words to their stems, which incorporates a great deal of language-dependent linguistic knowledge. Behind stemming, the hypothesis is that words with the same stem or word root mostly describe same or relatively close concepts in text and so words can be conflated by using stems. For example, the words, user, users, used, using all can be stemmed to the word 'USE'. In the present work, the Porter Stemmer algorithm [7], which is the most commonly used algorithm in English, is used.

D. Document Indexing

The main objective of document indexing is to increase the efficiency by extracting from the resulting document a selected set of terms to be used for indexing the document. Document indexing consists of choosing the appropriate set of keywords based on the whole corpus of documents, and assigning weights to those keywords for each particular document, thus transforming each document into a vector of keyword weights. The weight normally is related to the frequency of occurrence of the term in the document and the number of documents that use that term.

E. TermWeighting

In the vector space model, the documents are represented as vectors. Term weighting is an important concept which determines the success or failure of the classification system. Since different terms have different level of importance in a text, the term weight is associated with every term as an important indicator [8]. The three main components that affect the importance of a term in a document are the Term Frequency (TF) factor, Inverse Document Frequency (IDF) factor and Document length normalization [9].Term frequency of each word in a document (TF) is a weight which depends on the distribution of each word in documents. It expresses the Importance of the word in the document. Inverse document frequency of each word in the document database (IDF) is a weight which depends on the distribution of each word in the document database. It expresses the importance of each word in the document database [10]. TF/IDF is a technique which uses both TF and IDF to determine the weight a term. TF/IDF scheme is very popular in text classification field and almost all the other weighting schemes are variants of this scheme [9].In vector space model organization of document also affect the performance of system [11]. In this experiment we use term frequency method, other are also acceptable.

F. Feature Selection

The document term matrix contains the set of document as row and set of terms as columns. These terms are also known as features because there are used to uniquely identify the document. The sparsity of document term matrix represent the set of features that's frequency is zero. Higher the sparsity value lead to increase the set of feature and lower the sparsity value decrease the set of feature. Document frequency (DF) is the number of documents in which a term occurs. DF thresholding is the simplest technique for feature reduction. Stop word elimination explained previously, removes all high frequency words that are irrelevant to the classification task, while DF thresholding removes infrequent words. All words that occur in less than 'm' documents of the text collection are not considered as features, where 'm' is a pre-determined threshold. DF thresholding is based on the assumption that infrequent words are not informative for category prediction. DF thresholding easily scales to a very large corpora and has the advantage of easy implementation. In the present work, during classification, the document frequency threshold is set as sparsity of matrix that varies from 0.1 to 0.9.

III. EXPERIMENTAL SETUP

We have used Google Web API for aggregation of web data. We have aggregate 64 contents for concept "ipad". This experiment system is implemented by R including tm package. We have created one matrix as a document term matrix that has 9998 features. The Fig.2 shows the word cloud of used data set.

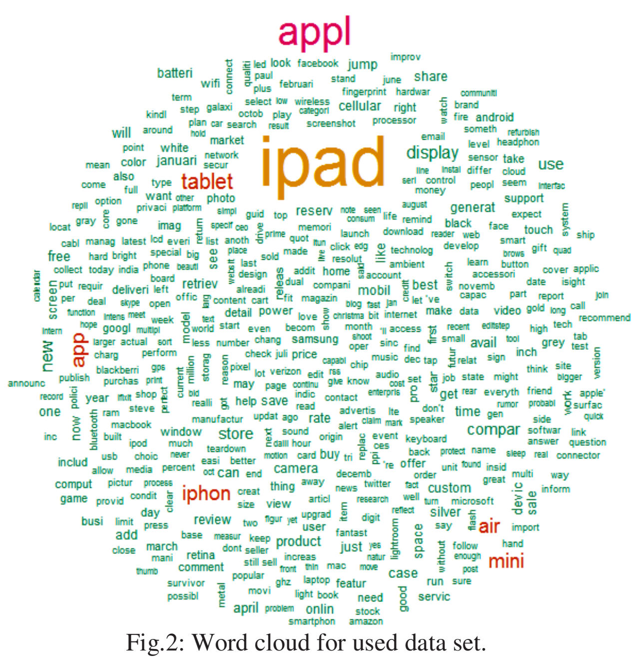

Fig.2: Word cloud for used data set.

IV. RESULT AND DISCUSSION

. The experiment was conducted with 64 documents, having 9998 unique terms. The experiments have been conducted using nine documents frequency threshold values (sparsity value in %), namely, 10, 20, 30, 40, 50, 60, 70, 80 and 90. The thresholding is the percentage value rather than the sparsity value. Table 1 and 2 show the result after applying the preprocessing technique namely stop word removal and stemming

TABLE 1: IMPACT OF STOP WORD WITH DIFFERENT SPARSITY VALUE ON FEATURE SET

Sparsity	WithStopWord	WithoutStopWord
0.1	9	1
0.2	10	1
0.3	18	6
0.4	39	22
0.5	71	46
0.6	121	89
0.7	215	170
0.8	442	374
0.9	1019	936
1	9998	9793

From the table 1 it is clear that the removal of stop-words decrease the size of feature set. We found the maximum decrement in feature set at sparsity value 0.9 as 90%.

TABLE 2: IMPACT OF STEMMING WITH DIFFERENT SPARSITY VALUE ON FEATURE SET

Sparsity	Without stemming	with Stemming
0.1	1	1
0.2	1	2
0.3	6	9
0.4	22	31
0.5	46	59
0.6	89	115
0.7	170	217
0.8	374	411
0.9	936	930
1	9793	7338

Table 2 shows the impact of stemming on feature set for different sparsity value. From the table 2 it is clear that the stemming process affect significantly to the size of feature set with different sparsity value. As we increase the sparsity value the size of feature set also increase. Only for sparsity value 0.9 the feature set decrease from 9793 to 936. Fig. 3, it could be seen that the application of stop word removal and stemming techniques have a positive impact on the number of terms selected. The results further reveal an important fact that stemming, even though is very important is not making only very negligible difference in terms of number of terms selected.

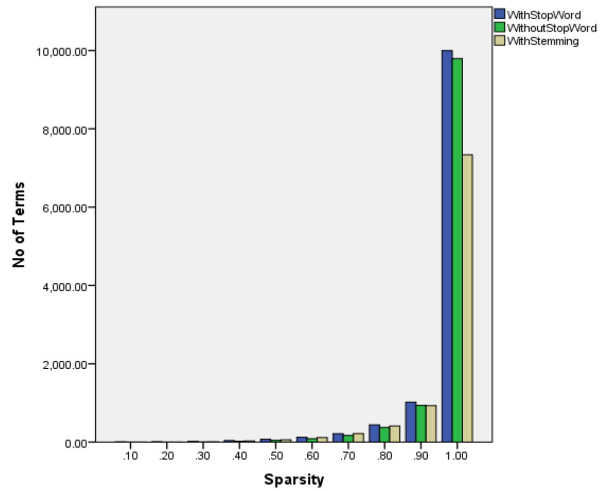

Fig.3: Comparison between techniques

V. CONCLUSION

The present work uses two important preprocessing techniques namely, stop word removal and stemming on web dataset. From the experimental results, it could be seen that preprocessing has a huge impact on performances of classification. The goal of preprocessing is to reduce the number of features which was successfully met by the selected techniques. From the results it is clear that the removal of stop-words decrease the size of feature set. For sparsity value 0.9 it decrease by 9%.On the other hand for stemming process As we increase the sparsity value the size of feature set also increase. Only for sparsity value 0.9 the feature set decrease from 9793 to 936.

REFERENCES

[1] K.Aas and A.Eikvil, "Text categorization: A survey", Technical report, Norwegian Computing Center, June, 1999.

[2] Katharina, M. and Martin, S. (2004) the Mining Mart Approach to Knowledge Discovery in Databases, Ning Zhong and Jiming Liu(editors), Intelligent Technologies for Information Analysis, Springer, Pp. 47-65.

[3] T. G. Kolda, D. P. O'Leary, "A semidiscrete matrix decomposition for latent semantic indexing information retrieval", Journal ACM Transactions on Information Systems (TOIS) TOIS Homepage archive vol.16(4), pp. 322-346, Oct. 1998.

[4] G.Salton, C. Buckley, "Term weighting approaches in automatic text retrieval," *Inf. Process. Manage*. 24, pp. 513–523, 1988.

[5] D. Harman, "Ranking algorithms. In Information Retrieval: Data Structures and Algorithms," *W. B. Frakes and R. Baeza-Yates, Eds. Prentice Hall, Englewood Cliffs*, NJ, pp.363–392, 1992.

[6] Xue, X. and Zhou, Z. (2009) Distributional Features for Text Categorization, IEEE Transactions on Knowledge and Data Engineering,Vol. 21, No. 3, Pp. 428-442.

ISROSET- Int. J. Sci. Res. in Computer Science & Engineering Vol-**3(2)**, PP (**1-4**) Apr 2015, E-ISSN: **2320-7639**

[7] Porter, M. (1980) An algorithm for suffix stripping, Program, Vol. 14, No. 3, Pp. 130–137.
[8] Karbasi, S. and Boughanem, M. (2006) Document length normalization using effective level of term frequency in large collections, Advances in Information Retrieval, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 3936/2006, Pp.72-83.
[9] Diao, Q. and Diao, H. (2000) Three Term Weighting and Classification Algorithms in Text Automatic Classification, The Fourth International Conference on High-Performance Computing in theAsia-Pacific Region,Vol. 2, P.629.
[10] Chisholm, E. and Kolda, T.F. (1998) New term weighting formulas for the vector space method in information retrieval, TechnicalReport, Oak Ridge National Laboratory.
[11]Sharma Dharmendra, jain suresh, "Content sharing in information storage and retrieval system using tree representation of documents",IEEE ,International conference on IT industry, business and government,CSIBIG2014 page 1-4,2014