ISSN:0975-9646

A Survey on Text Categorization Techniques for Indian Regional Languages

Pooja Bolaj[1], Sharvari Govilkar[2]
[1,2]*Department of Computer Engineering-PIIT,*
Mumbai University, India

***Abstract*-- The rapid development of the Information technology has led to the collection of documents in Indian regional languages. To classify millions of documents manually is an expensive and time consuming task. Therefore, automatic text classifiers are constructed which sort a given set of documents into different classes and whose accuracy and time efficiency is much better than manual text classification. This paper presents a survey of text categorization techniques for Indian regional languages**.

***Keywords*-- Text categorization, Clustering, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Hybrid Approach.**

I. INTRODUCTION

The development of Internet led to the exponential growth in the collection and availability of documents and managing such a huge collection of documents is difficult task. Therefore, automatic text categorization is used to categorize documents into various categories. Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set [4][7].

This paper is organized into 5 sections. The section 1 presents the introduction, section 2 describes about text categorization and its type**.** Related work is presented in section 3 which describes various text categorization techniques applied on Indian regional languages. Various text categorization techniques are discussed in section 4. Section 5and 6 offer comparison and observations of various categorization techniques applied on Indian regional. Conclusion is made in section 7.

II. TEXT CATEGORIZATION

Text categorization is an active research area of text mining to manage the information efficiently, by classifying the documents into classes using classification algorithms. Text categorization refers to solving the problem to classify documents based on their content into a certain number of predefined categories. The main aim of text categorization is to assign a category to a new document. The types of text categorization are as follows [8]:

A. Single-label Vs Multi-label text categorization

The case in which only one category is assigned to the input text is called single-label text categorization whereas the case in which more than one category can be assigned to the input text is called multi-label text categorization.

B. Category pivoted Vs Document pivoted text categorization

Category pivoted categorization is the process of assigning each document d ϵ D to a specific classifier c ϵ C. The alternative to this approach is document pivoted categorization in which we want to find every category c ϵ C under which a given document falls.

C. Hard categorization Vs Soft categorization

In hard categorization the classifier is required to firmly assign categories to document whereas in soft categorization the system ranks the various possible assignments and the final decision about class assignment is left to the user.

III. RELATED WORK

In this section we cite various text categorization techniques applied on different Indian regional languages to extract meaningful information and knowledge from unstructured text.

Jaydeep Jalindar Patil and Nagaraju Bogiri [1] provide automatic text categorization of Marathi documents based on the user's profile which includes user's browsing history. The system provides text categorization of Marathi documents by using the LINGO (Label Induction Grouping) algorithm. LINGO is based on VSM. The system uses the dataset containing 200 documents of 20 categories. The result represents that for Marathi text documents LINGO clustering algorithm is efficient.

Ashish Kumar Mandal and Rikta Sen [2] proposed how information from bangle online text documents can be categorized using four supervised learning algorithms, namely Decision Tree(C4.5), K-Nearest Neighbor(K-NN), Naive Bayes(NB), Support Vector Machine (SVM). The experimental results show that KNN and NB are more capable than SVM and Decision Tree (C4.5) in categorization of documents. Comparison of four classifiers in terms of training time indicates that all classifiers do not take the same learning time. Decision Tree (C4.5) takes more time than other three algorithms for training, whereas SVM is quick in learning.

Neha Dixit, Narayan Choudhary [3] proposed a rule based, knowledge –base driven tool to automatically classify Hindi verbs in syntactic perspective. They also provide of developing the largest lexical resource for Hindi verbs along with the information on their class based on valency and some syntactic diagnostic tests as well as their morphological/inflectional type.

ArunaDevi K., Saveetha R. [4] proposed an efficient method for extracting C-feature for classifying Tamil text documents. Using the C-feature extraction, we can easily classify the documents because C-feature will contain a pair of terms to classify a document to a predefined category.

Nidhi and Vishal Gupta [5] proposed an existing classification algorithm such as Naïve Bayes, Centroid based techniques for Punjabi Text Classification. And one new approach is proposed for the Punjabi Text Document which is the combination of Naïve Bayes and ontology based classification. The third derived approach is Hybrid approach which is a combination of Naïve Bayes and ontology based classification techniques. In this approach Naïve Bayes is used as Feature Extraction method for text classification and then ontology based classification algorithm is performed on extracted features. It is observed that Hybrid classification gives better result in comparison to Centroid based classifier and Naïve Bayes classifier that shows comparatively low results.

Nidhi, Vishal Gupta [6] introduced preprocessing techniques, features selection methods for Punjabi language and classification algorithm to classify the Punjabi text documents. The authors proposed domain based ontology algorithm for classification of Punjabi documents related to sports domain.

Nadimapalli V Ganapathi Raju et al. [7] have implemented the K-Nearest Neighbor (K-NN) algorithm, which is known to be one of the top performing classifiers applied for the English text. The results show that K-NN is applicable to Telugu text.

K. Rajan et. al. [8] presented text classification using Vector Space Model and Artificial Neural Network for morphologically rich Dravidian classical language Tamil. The experimental results show that Artificial Neural Network model achieves 93.33% on Tamil document classification.

Abbas Raza Ali, Maliha ljaz [9] compared statistical techniques for text classification using Naïve Bayes and Support Vector Machines, in context of Urdu language. Language specific preprocessing techniques are applied on it to generate standardized and reduced-feature lexicon.

Munirul Mansur et. al. [10] proposed n-gram based algorithm for Bangla text classification and to analyze the performance of the classifier Prothom-Alo news corpus is used. The result shows that as we increase the value of n from 1 to 3 performance of the text classification also increases, but from value 3 to 4 performance decreases.

Kavi Narayana Murthy [11] proposed supervised classification using the Naïve Bayes classifier has been applied to Telugu news articles in four major categories totaling to about 800 documents category-wise normalized tf-idf are used as feature values.

Meera Patil and Pravin Game [12] proposed an efficient Marathi text classification system using Naïve Bayes, Centroid, K-Nearest Classifier and Modified K Nearest Classifier. The authors also compared these classifiers for Marathi text documents and concluded that Naïve Bayes is the most efficient among the four considering classification accuracy and classification time.

IV. TEXT CATEGORIATION TECHNIQUES

Text categorization tasks can be divided into two sorts: supervised document classification, where some external mechanisms (such as human feedback) provide information on the correct classification for documents and unsupervised classification, where the classification must be done entirely without reference to external information. There is also semi-supervised document classification, where parts of the documents are labeled by the external mechanism.

A growing number of statistical categorization methods and machine learning approaches or more specifically supervised learning methods have been applied to document classification which includes Decision Tree, K-Nearest Neighbor, Neural Networks, Bayesian approaches (Naïve Bayes, non-Naïve Bayes), Vector based methods(Support Vector Machine and Centroid algorithm) etc. Several clustering techniques are also applied like K-means and Label Induction Grouping algorithm. The above techniques are briefly explained below:

A. Categorization Techniques

Different categorization techniques used to categorize documents are briefly explained below**:**

1) Decision Tree: Decision tree methods [18] reconstruct the manual categorization of the training documents in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. When the tree has created, a new document can simply be categorized by placing it in the root node of the tree and let it run through the query structure until it reaches a certain leaf.

2) K-Nearest Neighbor: KNN is a statistical approach [17][16] for text classification where objects are classified by voting several labeled training examples with their smallest distance from each object. The KNN classification method is outstanding with its simplicity and is widely used techniques for text categorization.

3) *Neural Network:* Neural network [17] is also called artificial neural network is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. Different neural network approaches have been applied to document categorization problems. While some of them use the simplest form of neural networks, known as perceptions, which consist only of an input and an output layer, others build more sophisticated neural networks with a hidden layer between the two others.

4) Naïve Bayes: A naïve Bayes classifier [16] is a simple probabilistic classifier based on applying Bayes theorem with strong independent assumptions. A naïve Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. Depending on the precise nature of the probabilistic model, naïve

Bayes classifiers can be trained very efficiently in a supervised learning setting.

5) *Vector Based Methods:* The two types of vector-based methods [17]: The centroid algorithm and support vector machines. From these two algorithms centroid is simplest.

a. Centroid Algorithm: During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category. This algorithm is also appropriate if number of categories is very large. Centroid algorithm computes similarity of test document with each centroid using cosine similarity measure [12]. It assigns a document, class with whose centroid a document has greatest similarity.

*b. Support Vector Machine (SVM) :*The main idea of SVM is to find a hyper-plane that best separates the documents and the margin, distance separating the border of subset and the nearest vector document, is large as possible. The nearest samples of the hyper-plane named support vectors are selected. The calculated hyper-plane permits to separate the space in two areas. To classify the new documents, calculate the area of the space and assign them the corresponding category.

B. Clustering Techniques

Clustering of documents [1] is mainly used to minimize the amount of text by categorizing or grouping similar data items. This grouping is common way for human processing information, and one of the good techniques for clustering helps to build different varieties which provide automated tools. The following is brief introduction to some of the clustering techniques:

1) K-means Algorithm: It is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data.

2) LINGO Algorithm: Lingo algorithm is based on vector space model [1]. First it extracts the user readable and frequent words/phrases from the input documents. Further by performing the Reduction of Original Term Document Matrix with Singular Value Decomposition (SVD) method to reduce the term document matrix, and then it find the labels of clusters and then assigns documents to that cluster labels based on the similarity value.

.

V. COMPARISON OF TEXT CATEGORIZATION TECHNIQUES

In this section we compare various text categorization techniques for Indian Regional Languages

TABLE I
COMPARISON OF VARIOUS TEXT CATEGORIZATION TECHNIQUES FOR INDIAN REGIONAL LANGUAGES

Sr. No	Name of the Technique	Implemented on Languages	Observations/ Remarks
I	**Classification Techniques**		
1	Decision Tree (DT)	Bangla	Four promising Supervised learning methods applied on Bangla news corpus to evaluate the capabilities of these methods. It is observed that DT takes more time than other text categorization techniques for training [2].
2	K-Nearest Neighbor (K-NN)	Bangla	K-NN is more capable on small and organized training sets compared to other supervised learning algorithms [2].
		Telugu	The value of k for K-NN (k=nearest training document) increases up to a certain limit maximum correct similar documents. If k value increases above this limit the proportionality of correct documents decrease. For K-NN classifier of Telugu documents the limit is 10 [4].
		Marathi	K-Nearest Classifier has least classification accuracy compared to Naïve Bayes, Centroid and Modified K-NN. Multi-core implementation of the classifiers reduces classification time for K-NN type classifier drastically [12].
3	Centroid Algorithm	Punjabi	Among the three methods NB, Centroid based and Hybrid applied on Punjabi documents Centroid achieved comparatively low result in terms of F-score [5].
4	Support Vector Machine (SVM)	Bangla	Average F-measure produces the best result for SVM followed by Naïve Bayes (NB), DT and K-NN. Also time taken by SVM for training is less compared to NB, DT, and K-NN [2].
		Urdu	Text categorization for Urdu language uses language specific preprocessing techniques to generate standardized and reduced-feature lexicon. NB is very efficient but SVM is more accurate for categorization of Urdu documents [9].
5	Naïve Bayes (NB)	Bangla	Naïve Bayes is more efficient in terms of training document sets. NB gives better results for F-measures after SVM [2].
		Punjabi	When all three NB, Centroid based and Hybrid algorithms are applied it is observed that more than NB Hybrid gives best results for Punjabi text categorization [5].

Sr. No	Name of the Technique	Implemented on Languages	Observations/ Remarks
		Urdu	The Urdu text classification using NB is very efficient. Ii also concluded that stemming algorithm used for this classification decreases the overall system accuracy [9].
		Telugu	NB is applied for news articles which in normalized TFXIDF are used to extract features from documents. And it is observed that the precision resulted is 93% [11].
		Marathi	Naive Bayes is efficient among four classifiers i.e. Naïve Bayes, Centroid, K-NN and Modified K-NN in terms of classification accuracy and classification time. Whereas K-NN has least accuracy among the four techniques [12].
6	Neural Networks(NN)	Tamil	It is observed that NN models are effective in representation and classifying Tamil documents. The performance of NN is better for more representative collection [8].
II	**Clustering Techniques**		
1	LINGO	Marathi	The LINGO algorithm based on Vector Space Model. The system automatically categorizes Marathi documents based on the User's profile inducing User's browsing history [1].
III	**Hybrid Approach**	Punjabi	It is observed that among Hybrid classification gives better results in comparison to Centroid Based classifier and Naïve Bayes classifier that show low results in terms of F-score. This is due to the reason that after the removal of stopwords, system cannot find important features that increase the classification rate [5].

VI. CONCLUSION

In this paper, we discussed the various techniques of text categorization for Indian regional languages. From literature survey it is observed that three supervised learning methods Support Vector Machine (SVM), Naïve Bayes (NB) and K-Nearest Neighbor (K-NN) are most suitable and give better results for document classification for Indian regional languages like Bangla, Telugu, Urdu, Punjabi, Marathi and Tamil. Clustering technique LINGO is better suited and only implemented technique for Marathi language.

ACKNOWLEDGMENT

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

REFERENCES

[1] Jaydeep Jalindar Patil, Nagaraju Bogiri, "Automatic Text Categorization Marathi Documents", 2321-7782, *International Journal of Advance Research in Computer Science and Management Studies,* March-2015.

[2] Ashis Kumar Mandal, Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization", *International Journal of Artificial Intelligence & Application (IJAIA),* DOI:10.5121/ijaia.2014.5508 September 2014.

[3] Neha Dixit, Narayan Choudhary, "Automatic Classification of Hindi Verbs in Syntactic Perspective", 2250-2459, *International Journal of Emerging Technology and Advanced Engineering,* August 2014.

[4] ArunaDevi, K., Saveetha, R., "A Novel Approach on Tamil Text Classification Using C-Feature", 2321-0613, 2014. *IJSRD-International Journal of Scientific Research & Development,* 2014.

[5] Nidhi, Vishal Gupta, "Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach", DOI: 10.5121/csit.2012.2421.

[6] Nidhi, Vishal Gupta, "Algorithm for Punjabi Text Classification", 0975-8887, *International Journal of Computer Applications,* January-2012.

[7] Nadimapalli V Ganapathi Raju et. al., "Automatic Information Collection & Text Classification for Telugu Corpus using K-NN", 2231-1009, *International Journal of Research in Computer Application & Management,* November-2011.

[8] K. Rajan et. al., "Automatic classification of Tamil documents using vector space model and artificial neural networks", *Expert Systems with Applications* 36 (2009) 1091-10918, ELSEVIER, 2009.

[9] Abbas Raza Ali, Maliha Ijaz, "Urdu Text Classification", FIT'09, December 16-18, 2009, CIIT, Abbottabad, Pakistan.

[10] Munirul Mansur, NaushadUzZaman , Mumit Khan, "Analysis of N-Gram Based Text Categorization for Bangla in Newspaper Corpus".

[11] Kavi Narayan Murthy, "Automatic Categorization of Telugu News Articles".

[12] Meera Patil, Pravin Game, "Comparison of Marathi Text Classifiers", *ACEEE Int. J. on Information Technology,* DOI: 01.IJIT.4.1.4, March 2014.

[13] István Pilászy, "Text Categorization and Support Vector Machines".

[14] Bijal Dalwadi, Vishal Polara, Chintan Mahant, "A Review: Text Categorization for Indian Language", 2349-4476, *International Journal of Engineering Technology, Management and Applied Sciences,* March 2015.

[15] Bhumika, Prof. Sukhjit Singh Sehra, Prof. Anand Nayyar, "A Review Paper on Algorithms Used for Text Categorization", 2319-4847, *International Journal of Application or Innovation in Engineering Technology & Management,* March 2013.

[16] B. Mahalakshmi, Dr. K. Duraiswamy, "An Overview of Categorization Techniques" 2249-6645, *International Journal of Modern Engineering Research (IJMER).* Oct 2012.

[17] S. Niharika, V. Sneha Latha, D. R. Lavanya, "A Survey on Text Categorization", 2231-2803, *International Journal of Computer Trends and Technology,* 2012.

[18] Meenakshi, Swati Singh, "Review Paper on Text Categorization Techniques" ISSN: 2348-8387, *SSRG International Journal of Computer Science and Engineering(SSRG-IJCSE)-EFES* April 2015.