Vol.4,No.1, July, 2015
162 - 167,ISSN:0975-5446 [VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]

A Study of Text Classification Natural Language Processing Algorithms for Indian Languages

Jasleen Kaur
[1]Assistant Professor
[2]Research Scholar
[1]Shroff S. R. Rotary Institute of Chemical Technology
[2]Uka Tarsadia University,
[1]Ankleshwar, Gujarat, India
[2]Bardoli, Gujarat, India
sidhurukku@yahoo.com

Dr.Jatinderkumar R. SAINI
[1]Associate Professor
[2]Research Supervisor
[1]Narmada College of Computer Application
[2]Uka Tarsadia University,
[1] Bharuch, Gujarat, India
[2]Bardoli, Gujarat, India
saini_expert@yahoo.com

Abstract

In this informative age, many documents in different Indian Languages are available in digital forms. For easy retrieval of these digitized documents, these documents must be classified into a class according to its content. Text Classification is an area of Text Mining which helps to overcome this challenge. Text Classification is act of assigning classes to documents. This paper provides the analysis of Text Classification works done on Indian Language content. Text being present in Indian language imposes the challenges of Natural Language processing. This study shows that supervised learning algorithms (Naive Bayes (NB), Support Vector Machine (SVM), Artificial Neural Network (ANN), and N-gram) performed better for Text Classification task.

Keywords: Classification, Naive Bayes, Natural Language Processing, Supervised, Support Vector Machine.

1. Introduction

With the advent of World Wide Web, amount of data on web increased tremendously. Although, such a huge accumulation of information is valuable and most of this information is texts, it becomes a problem or a challenge for humans to identify the most relevant information or knowledge. So text classification helps to overcome this challenge. Text classification is the act of dividing a set of input documents into two or more classes where each document can be said to belong to one or multiple classes [1]. Text Classification is a text mining technique which is used to classify the text documents into predefined classes. Classification can be manual or automated. Unlike manual classification, which consumes time and requires high accuracy, Automated Text Classification makes the classification process fast and more efficient since it automatically categorizes document.

Language is used as medium for written as well as spoken communication. With the use of Unicode encoding, text on web may be present in different languages. This will add complexity of natural language processing to text classification. Text Classification is combination of Text Mining as well as Natural Language Processing. It has many applications such as document indexing, document organization and hierarchical categorization of web pages [1]. This task is usually solved by combining Information Retrieval (IR) technology and Machine Learning (ML) technology which both work together to assign keywords to the documents and classify them into specific categories . ML helps to categorize the documents automatically and IR represents the text as a feature.

This paper concentrates on analysis of text classifier work on different Indian Languages. Section II brief about steps involved in text classification process. Section III discusses approaches used in text classification and work done in Indian Languages.

2. Text Classification Process

Text Classification process consists of various sub phases, each of which has its own importance and need, as shown in figure 1. Text Classifier, in general, consists of six basic sub parts: Data Collection, Pre-processing phase, Feature Extraction, feature selection, Building a classifier, performance evaluation [1] [2]. The purpose and use of each sub phase is mentioned below:

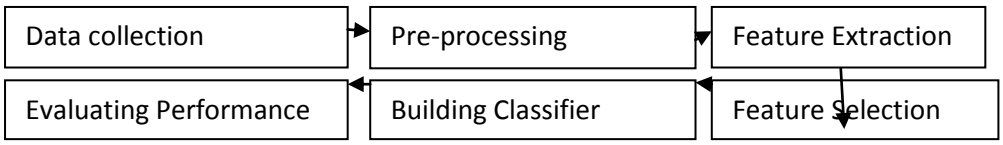

Figure 1: Basic Text Classification Process

2.1 Data Collection

This is first step of classification process is corpus building which consists of collecting the different types (format) of document like .html, .pdf, .doc, web content etc. These documents are used during training and testing the classifier.

2.2 Pre-Processing

The first step of pre-processing which is used to presents the text documents into clear word format. The documents prepared for next step in text classification are represented by a great amount of features. Commonly the steps taken are:
-*Tokenization:* A document is treated as a string, and then partitioned into a list of tokens.
-*Removing stop words*: Stop words such as "the", "a", "and", etc. are frequently occurring, so the insignificant words need to be removed.
-*Stemming word:* Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute.

2.3 Feature Extraction

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector. The most commonly used document representation is called vector space model (VSDM), documents are represented by vectors of words. Some of limitations are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document. To overcome these problems, term weighting methods can be used to assign appropriate weights to the term.

2.4 Feature Selection

After pre-processing and feature extraction the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space. Many feature evaluation metrics have been notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index [1].

2.5 Classification

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi-supervised methods [1]. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayes classifier, Decision Tree, K-nearest neighbour(KNN), Support Vector Machines(SVMs), Neural Networks, Rocchio's.

2.6 Performance Evaluation

The performance of Text Classification System can be evaluated by using four metrics: Accuracy, Precision, Recall and F1 measure. Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives.

$$Precision = \frac{No.of\ correct\ extracted\ text}{Total\ no.of\ extracted\ texts}$$

Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives.

$$Recall = \frac{No.of\ correct\ extracted\ texts}{Total\ number\ of\ annotated\ texts}$$

Precision and recall can be combined to produce a single metric known as F-measure, which is the weighted harmonic mean of precision and recall. The main advantage of using F-measure is it is able to rate a system with one unique rating.

$$F-measure = \frac{2 * precision * recall}{precision + recall}$$

Accuracy measures the overall degree to which instances have been correctly classified, using the formula as defined below

.

$$Accuracy = \frac{No.of\ correctly\ classified\ instances}{Total\ no.of\ instances}$$

3. Approaches for Text Classification Task

In relation to sentiment analysis, the literature survey done indicates two types of techniques – Supervised and Unsupervised learning.

3.1 Supervised learning

In a machine learning based classification, two sets of data are required: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of data, and a test set is used to validate the performance of the automatic classifier. A number of machine learning techniques have been adopted to classify the reviews. Machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) have achieved great success in text categorization. The other most well-known machine learning methods in the natural language processing area are K-Nearest neighbourhood, ID3, C5, centroid classifier, winnow classifier, and the N-gram model.

N**aive Bayes Classifier** is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions One of the main reasons that NB model works well for text domain because the evidences are "vocabularies" or "words" appearing in texts and the size of the vocabularies is typically in the range of thousands. The large size of evidences (or vocabularies) makes NB model work well for text classification problem. The Naive Bayes algorithm is widely used algorithm for document classification for content present in Indian Language [2][3][4][5][6].The support vector machine is a statistical classification method proposed by Vapnik [7].SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set .Support vector machines (SVM), a discriminative classifier is considered the best text classification method [1][6]. The K-nearest neighbour (KNN) is a typical example based classifier that does not build an explicit, declarative representation of the category, but relies on the category labels attached to the training documents similar to the test document. Given a test document d, the system finds the k nearest neighbours among training documents [1]. Rocchio's Algorithm [8] is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class, and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

In all supervised approaches, reasonably high accuracy can be obtained subject only to the requirement that test data should be similar to training data. To move a supervised text classifier to another domain would require collecting annotated data in the new domain and retraining the classifier. This dependency on annotated training data is one major shortcoming of all supervised methods.

3.2 Unsupervised Learning:

All approaches previously described build upon a set of fully annotated data, which is used to train a classifier with one technique or another. This classifier is then used to classify novel incoming text. Much of the research in unsupervised text classification makes use of lexical resources available [9].

3.3 Analysis of Text Classification works in Indian Languages

Indian Languages belongs to three language families: Indo-Aryan, Indo-Dravidian and Sino-Tibetan. Indo-Aryan consists of Punjabi, Bengali, Hindi, Urdu, Oriya; Indo-Dravidian consists of Telugu, Kannada, Tamil, Malayalam; and Sino-Tibetan consists of Manipuri, Meithei, and Himalayish [10]. This section provides the brief survey about work done in Text Classification problem in Indian Languages (as shown in Table 1).

Statistical techniques using Naïve Bayes and Support Vector Machine are used to classify subjective sentences from objective sentences for Urdu language, in this, language specific pre-processing used to extract the relevant features. As Urdu language is morphological rich language, this makes the classification task more difficult. The result of this implementation shows that accuracy, performance of Support Vector Machines is much better than Naïve Bayes Classification techniques [6].For Bangla

Text Classification, n-gram based algorithm is used and to analyze the performance of the classifier Prothom-Alo news corpus is used. The result show that with increase in value of n from 1 to 3, performance of the text classification also increases, but from value 3 to 4 performance decreases [11].

Naïve Bayes Classifier has been applied to Telugu news articles to classify 800 documents into four major classes. In this, normalized term frequency- inverse document frequency is used to extract the features from the document. Without any stop word removal and morphological analysis, at the threshold of 0.03, the classifier gives 93% precision [4].For morphologically rich Dravidian classical language Tamil, text classification is done using Vector Space Model and Artificial Neural network. The experimental results show that Artificial Neural network model achieves 93.33% which is better than the performance of Vector Space Model which yields 90.33% on Tamil document classification [12]. A new technique called Sentence level Classification is done for Kannada language; in this sentences are analyzed to classify the Kannada documents as most user's comments, queries, opinions etc are expressed using sentences. This Technique extended further to sentiment classification, Question Answering, Text Summarization and also for customer reviews in Kannada Blogs [5].
Very few works in literature are found in field of Text Classification in Punjabi Language. Domain Based text classification is done by Nidhi and Vishal G. [2].This classification is done on sports category only. In this work, two new algorithms, Ontology Based Classification and Hybrid Approach are proposed for Punjabi Text Classification. The experimental results conclude that Ontology Based Classification (85%) and Hybrid Approach (85%) provide better results.

Sarmah, Saharia and Sarma [13] presented an approach for classification of Assamese documents using Assamese WordNet. This approach has accuracy of 90.27 % on Assamese documents. Frequent terms of Assamese document are searched in Assamese WordNet. For each frequent term sysnet is found in WordNet, extended form of it is found and is associated with it. It searches each pre-defined class for extended terms present in testing document. It assigns the test document a class with which it has highest number of matching terms.

Table 1. Analysis of Text Classification in Indian Langauges

S.No.	Language Family	Languages	Author	Technique Used	Performance
1.	Indo-Aryan	Urdu	Ali R. A.[6]	Naive Bayes, SVM	SVM outperforms NB
		Bangla	Mansur M[11]	N-gram using normalised frequency	n-gram with value 2 or 3 are more useful for text classification
		Punjabi	Nidhi[2]	Naïve Bayes Ontology Based Hybrid	85%(Hybrid)
2.	Indo Dravidian	Telugu	Murthy K.N [4]	Naïve Bayes	93%
		Tamil	Rajan[12]	Artificial Neural Network	93.3%
		Kannada	Jayashree [5]	Naïve Bayes	0.67(average precision)
3.	Sino-Tibeto	Assamese	Sarmah[13]	Using Assamese wordnet	90.27%

4. Conclusion

Text Classification plays an important role in an area of Text Mining. Because of large of amount of data availability on web, for its easy retrieval this data must be organised according to its content.

Expansion of social media leds to usage of different kinds of languages on web. This adds complexity to Text classification task. India, being a unity in diversity, consists of many languages used for verbal and written communication. Use of Indian Languages gained popularity from last decade. This paper provides the analysis of various text classifiers which work on Indian Languages. Very few works are found in the field of Text Classification in Indian Languages. This study shows the supervised approaches worked well on Indian Language text classification but Indian content still need to be explored in terms of text classifications**.**

References

[1]Sebastiani, F., Machine learning in automated text categorization, ACM Computing Surveys, 34(2002), 1-47.

[2] Nidhi, Gupta V., Domain Based Classification Punjabi Text Documents, Proceedings of COLING 2012: Demonstration Papers, 2012, 297–304.

[3] Zheng G.,Tian Y.,Chinese web text classification system model based on Naive Bayes, International Conference on E-Product E-Service and E-Entertainment (ICEEE),2010,1-4.

[4]Murthy K.N.,Automatic Categorization of Telugu News Articles, Department of Computer and Information Sciences, University of Hyderabad, Hyderabad,2003, DOI= 202.41.85.68.

[5] Jayashree, R., An analysis of sentence level text classification for the Kannada language, International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2011, 147-151.

[6] Ali R. A. and Maliha I., Urdu Text Classification, Proceedings of the 7th International Conference on Frontiers of Information Technology, ACM New York, USA, 2009 ISBN: 978-1-60558-642-7 DOI= 10.1145/1838002.1838025.

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

[8] Rocchio, J, Relevance Feedback in Information Retrieval, In G. Salton (ed.). The SMART System,67-88.

[9] Youngjoong Ko Jungyun Seo, Automatic Text Categorization by Unsupervised Learning, Proceedings of the 18th conference on Computational linguistics, 1(2000), 453-459.

[10] Kaur J. and Saini JR, A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families, International Journal of Data Mining and Emerging Technologies, ISSN 2249-3220, 4(2),2014,53-60.

[11] Mansur M., UzZaman N., Khan M., Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus, Proceedings of 9th International Conference on Computer and Information Technology, Dhaka, Bangladesh. 2006.

[12] Rajan, K., Ramalingam, V., Ganesan, M., Palanive, S., Palaniappan, B., Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network, Expert Systems with Applications, Elsevier, 36(8),2009,10914–10918.

[13] J Sarmah, N Saharia, and S.K Sarma, A Novel Approach for Document Classification using Assamese WordNet,6th International Global Wordnet Conference, 2012, 324-329.