A Review of Document Classifier Systems Developed for Indian Languages

Harshali B. Patil
Department of Computer Science Engineering
Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya, Jalgaon, India

Abstract— In this digital era, massive amount of e-data is available in various languages and in several forms: like text, images, audio, video, etc. Automatic management of this data is a major challenge for computer science, which can be solved by using automatic classification systems. Automatic text classification process classifies text documents into predefined classes. This paper reviews the automated document classifier systems available for Indian languages. In this paper several document classifiers, along with document classification techniques and evaluation measures are discussed.

Keywords: Document Classifier Systems, Indian Languages

I. INTRODUCTION

Due to the proliferation of the Internet massive amount of e-data is available in various forms like text documents, images, audio data, video data, etc. Text documents are one of the richest sources of data for instance technical documents, user reviews, news articles, emails, etc. they all contain valuable information which can be used to find valuable insights from these data. During last two decades there is massive growth in number of digital text documents, hence researchers are attracted towards the development of systems that will automatically organize and classify these documents. Research into text classification aims to partition unstructured sets of documents into groups that describe the contents of the documents.

Document classification used in many applications like: e-mail filtering, spam filtering, mail routing, news monitoring, automated indexing of scientific articles, automated population of hierarchical catalogues of Web resources, identification of document genre, authorship attribution, survey coding and so on. An automated text classification is attractive because manually organizing huge set of text documents can be expensive, time-consuming and tedious task. Many Businesses are trying to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes. Text classification is the process of assigning tags or categories to the text according to its content. It's one of the fundamental tasks in Natural Language Processing (NLP) with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. Following figure depicts the document classification problem.

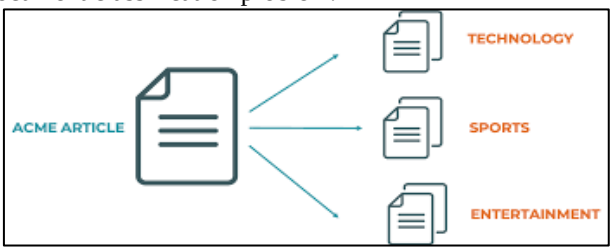


Fig. 1: Document classification

The text classification process can include different levels of scope: sub-sentence level, sentence level, paragraph level, and document level.

II. LITERATURE REVIEW

Document classification is one of the important research problems in the field of natural language processing. In literature, many document classifier systems are found for classifying English language documents. This paper reviews the documents classifiers systems developed for various Indian languages.

Abbas Ruza Ali et al. in 2009 compared statistical classifiers namely Naïve Bayes (NB) and Support Vector Machine (SVM) to classify Urdu documents. Relevant features are extracted by using language specific preprocessing steps and the experimental results show that SVM gives better result in terms of accuracy as compared to Naïve Bayes [8]. In 2009 K. Rajan et al presented classification of morphologically rich Dravidian language Tamil documents using Vector Space Model and Artificial Neural Networks. The result analysis shows that artificial neural network model achieves better performance (93.33%) than VSM (90.33%) on Tamil text documents [7]. Kavi Narayana Murthy [10] proposed automatic text classification with special emphasis on Telugu and concludes that the results obtained are comparable to published results for other languages. In 2011 Nadimapalli V Ganapathi Raju et.al 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 and their results show that K-NN is applicable to Telugu text [19]. ArunaDevi K. and Saveetha R proposed A Novel Approach on Tamil Text Classification Using C-Feature in 2014 and conclude that the documents can be easily classify using c-feature extraction because C-feature will contain a pair of terms to classify a document to a predefined category [18]. Ashis Kumar Mandal and Rikta Sen in 2014 described the classification of Web documents in Bangla language using four promising Supervised Learning algorithms: Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Byes (NB), and Support Vector Machines (SVM). News corpus from various Bangla websites is employed to evaluate the capabilities of these methods and concludes that, all four methods produce satisfactory performance. SVM attaining good result in terms of high dimensional and relatively noisy document feature vectors. [6]. Nidhi and Vishal Gupta described three classification techniques for Punjabi Text Classification that are Naïve Bayes, Centroid based and one new approach called Hybrid approach which is the combination of Naïve Bayes, ontology based. Among three techniques hybrid approach performs well [12]. Nidhi and Vishal Gupta [9] proposed domain based ontology algorithm for classification of Punjabi documents related to sports domain. Neha Dixit and Narayan Choudhary in 2014

proposed a rule based, knowledge–base driven tool to automatically classify Hindi verbs in syntactic perspective and reports an accuracy of more than 99% quantitatively and 100% using two specific corpora qualitatively. [21]. In 2019, Shalini Puri and Satya Prakash Singh developed An Efficient Hindi Text Classification Model Using SVM and reports 100% accuracy when the experiments have been performed on a set of four Hindi documents of two categories [22]. Sushma R. Vispute and M. A. Potey in 2020 proposed a system that contributes text classification for Marathi documents by using Vector Space Model based LINGO algorithm and obtained the overall accuracy of 91.10% [5].

From the literature survey it is observed that very few automated text document classification systems are developed for several Indian languages like: Tamil, Telugu, Bangala, Punjabi, Hindi & Marathi. It is also observed that only few techniques like: Naïve Bayes (NB), Support Vector Machine (SVM), VSM, K-Nearest Neighbor (KNN), ANN, Decision tree, domain-based ontology algorithm, rule-based and knowledge-based algorithm, VSM based LINGO algorithm are used for development of classifiers for Indian languages.

III. DOCUMENT CLASSIFICATION PROCESS

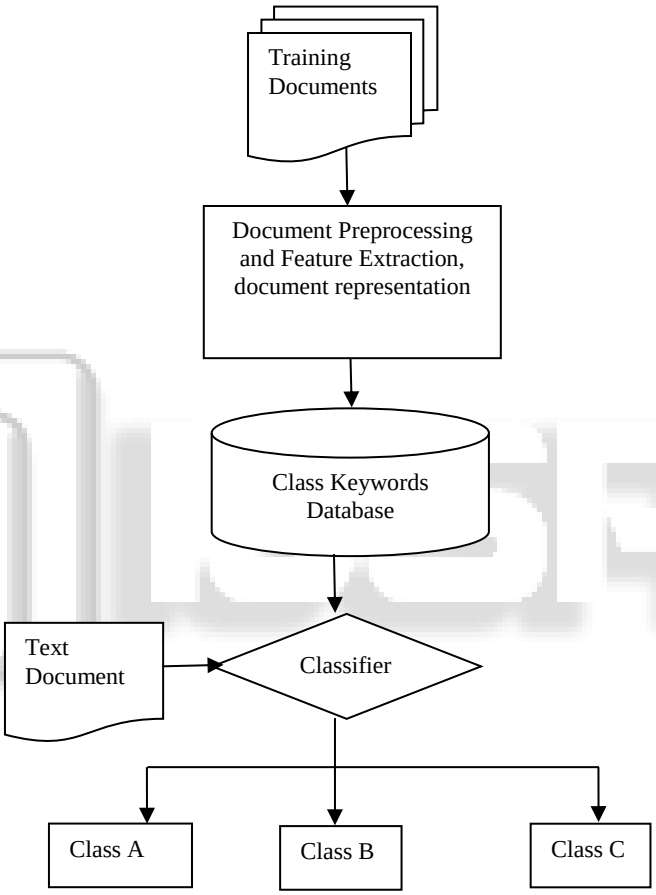


Fig. 2: Document Classification Process

The document classifier system consists of three main phases that includes: 1) Preprocessing 2) Feature Extraction and 3) Supervised / unsupervised learning Methods and produce the output as classified text documents. Pre-processing is one of the important phase of document classifier systems which consist of: Input Validation, Tokenization, Stop word Removal, Stemming, Morphological Analysis as the subphases. During input validation, the input document is validated to Devanagari script. If the input document consists of some terms, phrases or sentences in other languages or scripts then the non-Devanagari words/phrases are removed for further processing. Tokenization is the process of separating token from input text document. Stop-word removal removes the most frequently appearing words in a documents; which are insignificant for the processing of document. Stemming removes suffixes of word to get stem of word while Morphological Analysis is expected to produce root words for a given input document. The next step is feature extraction: during this phase the feature vector is computed. During this phase the frequent phrases or terms present in documents are obtained. In the last phase the classifier uses any supervised or unsupervised classification methods like: NB, KNN, SVM, decision tree, etc. to classify the input documents into their respective classes.

IV. DOCUMENT CLASSIFICATION TECHNIQUES

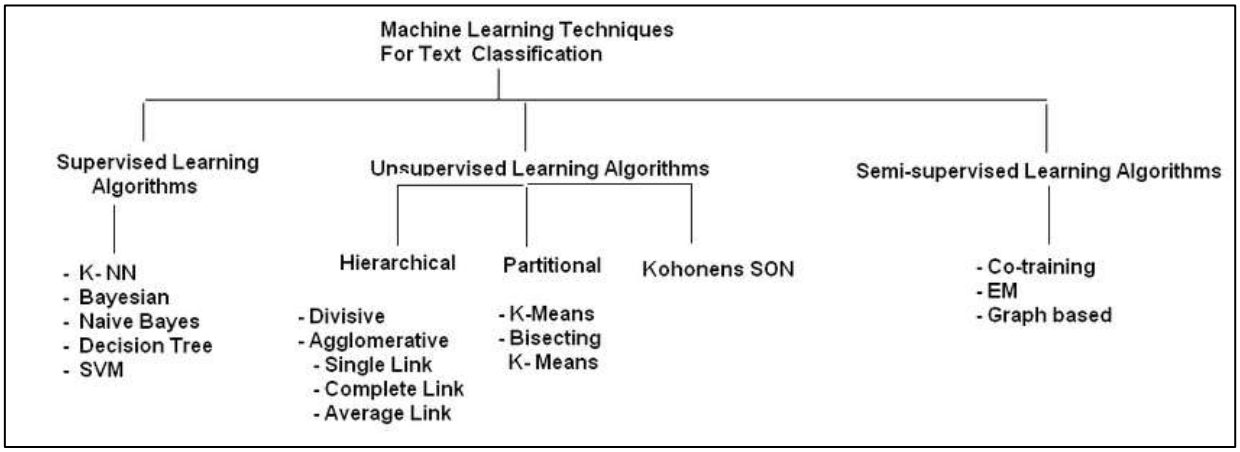


Fig. 3: classification Hierarchy of ML algorithms for text classification [4]

Fig. 3 shows several machine learning algorithms use for document classifications which includes: supervised, semi-supervised and unsupervised algorithms. This section discusses most popular supervised algorithms used for development of automated document classifiers of Indian languages.

SrNo	Technique/Algorithm	Language & Reference	Description
1	Naïve Bayes (NB)	Telugu[10], Urdu[8], Bangla[6], Punjabi [12]	Based on Bayes Theorem Simple probabilistic classifier. Assumes all of the features are mutually independent.
2	K-Nearest Neighbor (KNN)	Telugu [19], Bangla [6]	Instance based algorithm for text categorization. Document classification is done based on the majority of its nearest neighbors.
3	Support Vector Machine (SVM)	Urdu[8], Bangla [6], Hindi[22]	Based on statistical learning theory. searches for a decision boundary which separates the training data points into two classes known as support vector. chooses decision boundaries with largest margin.
4	Decision Tree	Bangla [6]	Inductive learning method, two steps involved: building and applying tree to database. Weights of the terms in the document are used. The leaf node label is then allocated the testing document.
5	Vector Space Model	Tamil [7] Marathi[5]	Document is represented in terms of term wight matrix. Euclidian distance or cosine similarity is use to measure the similarity between the documents.
6	Neural Network	Tamil [7]	The input units are assigned by its term weights, through the network, these units activation is propagated forward and the categorization decision(s) is determined by the value that the output unit(s) takes up.

Table I: Supervised Techniques Used For Development of Indian Document Classifier Systems

V. EVALUATION MEASURE FOR DOCUMENT CLASSIFICATION SYSTEM

Some of the measure used for evaluating document classifier systems are Classification accuracy, Confusion matrix, Precision and recall, F1 score [17].

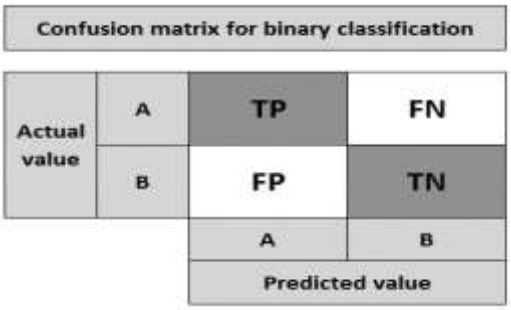
A. Classification Accuracy

Classification accuracy shows how many of the predictions are correct. Following formula is used to calculate the classifier accuracy :

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Number\ of\ all\ Predictions}$$

B. Confusion Matrix

Confusion matrix performs deeper analysis as compare to classification accuracy, because it shows the correct and incorrect (i.e. true or false) predictions of each class. In case of a binary classification task, a confusion matrix is a 2x2 matrix. If there are n different classes, then the confusion matrix will be of nxn size.



Here TP indicates True Positive means Positive class is predicted as Positive, FP indicates False Positive means Negative class is predicted as Positive, FN indicates False Negative means positive class is predicted as Negative, and TN indicates True Negative means Negative class is predicted as Negative.

C. Precision and Recall

Precision and recall metrics take the classification accuracy one step further and allow us to get a more specific understanding of document classifier model evaluation.
Precision measures how good our model is when the prediction is positive and it is calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

Precision indicates how many positive predictions are true while Recall measures how good our model is at correctly predicting positive classes.

$$Recall = \frac{TP}{TP + FN}$$

Recall indicates how many of the positive classes the model is able to predict correctly.

There is another measure that combines precision and recall into a single number, called as F1 score.

D. F1 Score

F1 score is the weighted average of precision and recall, which is calculated as follows:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 score is a more useful measure than accuracy for problems which includes uneven class distribution because it takes into account both the false positive and false negatives cases.

VI. Conclusion & Future Work

This paper reviews the document classification systems developed for Indian languages. During literature review, it is found that many document classification systems are developed for Tamil, Punjabi, Hindi, languages while few efforts related to document classification of Marathi language has also been reported. It is also observed that SVM, KNN and its variants, Naïve Bayesian algorithms are generally used for document classification of Indian languages. As far as Marathi is considered significant work has not been carried out related to document classification. Hence it would be interesting to develop an automated Marathi Document classifiers with the help of several techniques and compare the performance of these classifiers for Marathi document classification process.

References

[1] Pooja Bolaj and Sharvari Govilkar. Text Classification for Marathi Documents using Supervised Learning Methods. *International Journal of Computer Applications* 155(8):6-10, December 2016.
[2] Aishwarya Sahani, Kaustubh Sarang, Sushmita Umredkar, and Mihir Patil Automatic Text Categorization of Marathi Language Documents , International Journal of Computer Science and Information Technologies, Vol. 7 (5) , 2016, 2297-2301
[3] Jaydeep Jalindar Patil, Nagaraju Bogiri, "Automatic Text Categorization Marathi Documents", International Journal of Advance Research in Computer Science and Management Studies,Volume 3, Issue 3, March 2015 pg. 280- 287, http://ijarcsms.com/docs/paper/volume3/issue3/V3I3-0088.pdf
[4] Shweta Dharmadhikari, Maya Ingle, Parag Kulkarni : "Empirical Studies On Machine Learning Based Text Classification Algorithms", Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011, pp 161-169.
[5] Sushma R. Vispute, M. A. Potey, "Automatic Text Categorization of Marathi Documents Using Clustering Technique", 978-1-4673-2818-0/13, 2013 IEEE.
[6] Ashis Kumar Mandal, Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization", International Journal of Artificial Intelligence and Application (IJAIA), DOI: 10.5121/ijaia.2014.5508 September 2014.
[7] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan. 2009. " Automatic classification of Tamil documents using vector space model and artificial neural network". Expert Syst. Appl. 36, 8 (October, 2009), 10914–10918. https://doi.org/10.1016/j.eswa.2009.02.010
[8] Abbas Raza Ali, & Maliha Ijaz, "Urdu Text Classification", FIT'09, December 16-18, 2009, CIIT, Abbottabad, Pakistan. FIT '09: Proceedings of the 7th International Conference on Frontiers of Information Technology December 2009 Article No.: 21, Pages 1–7 https://doi.org/10.1145/1838002.1838025
[9] Nidhi, Vishal Gupta, "Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach", Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 109-122, COLING 2012, Mumbai, December 2012.
[10] Kavi Narayan Murthy, "Automatic Categorization of Telugu News Articles".
[11] A. Kanaka Durga, A. Govardhan, "Ontology Based Text Categorization-Telugu Documents", International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011. ISSN 2229-5518.
[12] Nidhi, Vishal Gupta, "Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach", DOI: 10.5121/csit.2012.2421.
[13] Vishnu Murthy.G, Dr. B. Vishnu Vardhan, K. Sarangam and P. Vijay pal Reddy, "A Comparative Study on Term Weighting Methods For Automated Telugu Text

Categorization With Effective Classifiers", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.6, November 2013.
[14] Vishal Polara, Bijal Dalwadi, Chintan Mahant "A Review: Text Categorization for Indian Language", 2349-4476, International Journal of Engineering Technology Management and Applied Sciences, March 2015.
[15] B. Mahalakshmi, Dr. K. Duraiswamy, "An Overview of Categorization Techniques", 2249-6645, International Journal of Modern Engineering Research (IJMER). October 2012.
[16] S.Niharika, V.Sneha Latha, D.R.Lavanya, "A Survey on Text Categorization", 2231-2803, International Journal of Computer Trends and Technology, 2012.
[17] https://towardsdatascience.com/how-to-best-evaluate-a-classification-model-2edb12bcc587
[18] ArunaDevi, K., Saveeth, R. "A Novel Approach on Tamil Text Classification Using C-Feature" IJSRD - International Journal for Scientific Research & Development, Vol. 2, Issue 05, 2014
[19] Nadimapalli V Ganapathi Raju, Vidya Rani V, Bhavya Sukavasi & Sai Rama Krishna Chava , "Automatic Information Collection & Text Classification for Telugu Corpus Using K-Nn Algorithm", Int. Journal of Research in Computer Application and Management, Volume No. 1 (2011), Issue No. 9 (November), pp88-93
[20] N. Dangre, A. Bodke, A. Date, S. Rungta, S.S. Pathak, System for Marathi News Clustering, Procedia Computer Science, Volume 92, 2016, Pages 18-22, ISSN 1877-0509.
[21] Neha Dixit, Narayan Choudhary: " Automatic Classification of Hindi Verbs in Syntactic Perspective" , International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 8, August 2014, pp 572-579
[22] Shalini Puri and Satya Prakash Singh, "An Efficient Hindi Text Classification Model Using SVM, S.-L. Peng et al. (eds.), Computing and Network Sustainability, Lecture Notes in Networks and Systems 75

