RESEARCH ARTICLE OPEN ACCESS

# A Review on Document Classifier System for Indian Languages

**Ms. Madhuri P. Narkhede [1], Dr. Harshali B Patil [2]**

[1] Department of Computer Science and information technology Department, Smt. Devkiba Mohansinhji Chauhan College of Commerce & Sciences, Silvassa, UT of DD & DNH-396230, India.
[2] Department of Computer Science, Dr. Annasaheb G. D. Bendale Mahila Mahavidyalaya, Jalgaon, Maharashtra, India.

## ABSTRACT

In this era of digitalization, digital information in the form of text documents such as news, social media chats, comments, company reports, reviews on products, medical reports, tweets, and so on is increasing rapidly. Since numerous electronic documents are available in various languages, it is necessary to classify them and extract meaningful information. Classifying these electronic documents manually is a very time-consuming and tedious task. Automated text classifier plays a crucial role in classifying these digital documents. This paper discusses various document classifier systems developed for Indian languages using machine learning techniques.

## I. INTRODUCTION

The tremendous growth in computers and internet users causes the generation of massive amount of digital documents. India is a diverse country where diversity is found everywhere, from religions to the cultures and languages the people speaks. India has 22 official languages, and massive amounts of critical textual data are available for these languages. Searching this massive data manually is a time-consuming and tedious task. An automated document classifier solves this problem. Document classification is useful to improve the performance of information retrieval systems. Efficient document classification is helpful for various Platforms, such as E-commerce, news agencies, content curators, blogs, directories, and user likes.

Marathi is the official language of Maharashtra state; hence, many important documents are available in Marathi. Hence, automated processing of these documents is a need of time. Several text document classifiers are available for many languages. However, more work is needed for Marathi Document classification [1]. This paper presents a detailed literature survey of document classifier systems available for several languages. The article has been outlined as follows: Section II discusses available literature, Section III discusses the Document Classification process, Section IV discusses various document classification Techniques, and Section V concludes the present review article.

## II. LITERATURE REVIEW

This literature review discusses different supervised learning techniques implemented and their results in various domains for classifying Indian and Foreign languages.

Naïve Bayes classifier is used to classify Telugu news articles using normalized TF-IDF to extract features from the document without stop word removal and stemming with 93% precision. The author concludes that Morphological analysis and stemming can improve performance [2]. NB and SVM are used to classify Urdu documents using language-specific pre-processing steps. The experimental results show that SVM gives better accuracy than Naive Bayes [3]. SVM and ANN are used for the classification of Tamil documents. It is analyzed that the SVM method needs a high dimensional space to represent the documents, and the ANN classifier requires fewer features. The result analysis shows that the artificial neural network model achieves better performance (93.33%) than SVM (90.33%) on Tamil text documents [4].

The ontology-based classifier is developed for the classification of Panjabi sports documents. The advantage of this ontology-based classification is that we do not need Training Data, i.e., labelled documents, to classify the documents. In contrast, other classification techniques, such as the KNN technique, Naïve Bayes algorithm, Association Classification, etc., need training sets or labelled documents to train the classifier to classify the unlabelled documents. The results show that these approaches provide better results than standard algorithms such as the Naïve Bayes classifier (NB) and Centroid classifier [6,7].

Classification methods such as decision trees, rules, Bayes, nearest neighbour classifiers, SVM classifiers, and neural networks have been extended for the generic text classification process. However, the author concludes that more feature selection improvements can still be made [8]. Named Entity Recognition is implemented to identify and classify Nepali text into predefined categories using a Support Vector Machine (SVM) on a small training dataset. The analysis shows that a large dataset can improve the classifier's performance [9].

It is also analyzed that Logistic regression and SVM perform better on sentence classification where the sentence is represented as a Bag of Words (BoW). It is observed that these models achieve excellent performance in practice. However, they require more time to train and test enormous datasets [10].

SVM is used for categorizing Bengali documents and achieved good results compared to a Decision tree, KNN, and NB [11]. SVM is also used for Hindi Text classification for small datasets and achieved 100% results [12]. Multinomial naive Bayes(MNB), support vector machine(SVM), and logistic regression algorithms are used for classification of Telugu text document classification using various feature extraction techniques like uni-gram and Bi-gram the author obtained 99% accuracy using SVM [13]. Deep learning classifiers such as CNN, RNN, and RNN with LSTM are used for text classification and analysis.

Though several classifiers are available in the literature, only a few worked on classifying Indian languages, especially the Marathi language [14].

An automatic convolutional neural network (CNN)-based method used to classify poems written in the Marathi language based on emotions. Experiments are performed with different models of CNN, considering different batch sizes, filter sizes, and regularization methods like dropout and early stopping conducted to analyze the performance. The result analysis shows that the proposed CNN architecture for the classification of poems produces an impressive accuracy of 73% [15].

A hybrid approach is developed by combining text-based and graph-based features for classification. The author obtained 98.73% accuracy using Naïve Bayes Multinomial (NBM) [16]. Text classification using a Bag-of-Words representation model with term frequency-inverse document frequency (TF-IDF) and word embedding technique 'GloVe' is implemented to find words with similar semantic meaning on three different datasets: BBC, Classic4, and 20-newsgroup. The performance of the proposed method is compared with other methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Latent Semantic Indexing (LSI), a hybrid approach PCA+LDA using the Naïve Bayes classifier and observed that the proposed algorithm gives better classification results than other dimension reduction techniques [17].

During the literature review, it was observed that though several text classification models are available for classifying Indian and Foreign languages It is observed from the literature review that many classifiers are available for foreign languages; however, in the context of Indian languages, much work needs to be done.

## III. DOCUMENT CLASSIFICATION PROCESS

Assigning a document to one or more categories based on its content is known as document classification. This procedure can be carried out manually or automatically, utilizing a variety of methods. The manual process can be time-consuming, error-prone, and costly. Automatic classification can improve the efficiency and accuracy of document management. Without human assistance, papers are categorized by Deep Learning algorithms.

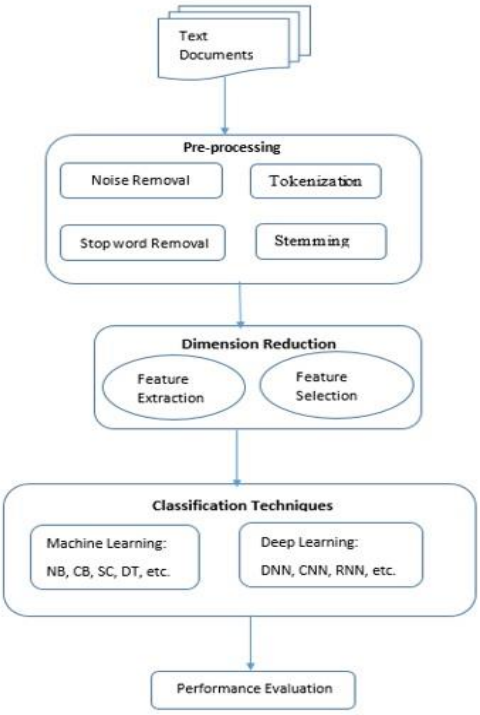

Fig. 1: General architecture of Document classifier system

Text classification comprises four levels: Document Level, Paragraph Level, Sentence Level, and Sub-Sentence Level. Classification of text undergoes various phases:

A. *STEP-1: Data Collection*

Data can be collected from various sources as per the research domain, and customized datasets can be developed for categories where datasets are unavailable.

B. *STEP-2: Preprocessing*

1) Noise removal: Input Validation is the first step in the preprocessing phase. The extra space, digits, special symbols, punctuation, and other language data are removed during this phase.

2) Tokenization: A whole document must be partitioned into a list of tokens for further processing. Tokenization converts documents into a set of tokens.

3) Stop word removal: Stop words occur frequently in the document but do not contribute to the meaning of the document; hence, to improve the accuracy of the automated systems, these words need to be omitted. Stop word removal removes these words.

4) Stemming: The process of reducing words to their root form is called stemming. Two forms of the same word cannot be distinguished and treated as completely different. Reducing the word to its root form can solve this problem.

C. *STEP-3: Feature Extraction & Selection:*

The term Frequency-Inverse Document Frequency (TF-IDF) Model, Word2Vec, and Global Vectors for Word Representation (GloVe) [17] can be used to extract features from the dataset.

Feature extraction will be used to select the required features to improve the performance of the classification model.

D. *STEP-4: Dimension Reduction*

To improve the efficiency of the classifier, data represented in high-dimension space can be reduced [18].

E. *STEP-5: Classification*

Document classification algorithms are classified into three categories: Supervised, Semi-supervised, and unsupervised. Supervised learning is a expensive classification technique that requires human intervention to assign labels to the data. Various Supervised learning algorithms will be trained and tested for Automated Marathi document classification. Different machine learning algorithms like Naïve Bayes, Support Vector Machine (SVM), and K-Nearest neighbor can be implemented to classify Marathi documents. Performance of the classification can be enhanced by training and testing various Deep learning frameworks like Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and RNN.

F. *STEP-6: Performance Evaluation*

Evaluating the model's performance is an essential element of any machine-learning workflow. Different performance metrics and techniques can be used to evaluate a classifier, like Accuracy, confusion matrix, precision, recall, f-measure, etc. The detailed evaluation measures are discussed in the literature [18].

## IV. DOCUMENT CLASSIFICATION TECHNIQUES

A. *Supervised document classification*

Supervised techniques require training corpora. In supervised technique the document category is assigned to the document based on the training data collection. From a conceptual standpoint, supervised approaches look at labeled historical data in an effort to determine a relationship between the document and its category.

1)*Naïve Bayes (NB):*

NB is the probabilistic classifier based on Bayes' theorem. The NB algorithm is suitable for Binary(two class) as well as Multiclass problems. The NB algorithm works well in diverse real-world problems such as text and document classification. Naïve Bayes is one of the most popular algorithms used for text classification. It requires a small training dataset representing the text as BoW. It is easy to understand and provides accuracy if appropriately trained [19].

2)*Support Vector Machine (SVM):*

SVM creates a hyperline for the classification of text data. It plots each data item as a point in n-dimensional space. A hyperline will be drawn to classify the data into predefined classes or categories. It is suitable to work on small datasets as it requires more time to train the model; the Noisy dataset may degrade the classifier's performance [19].

3)*K-Nearest Neighbor (KNN):*

KNN tests the degree of similarity between documents and k training data to determine the category of test documents. It is challenging to find K nearest neighbour from a large dataset [19].

*4) Decision Tree:*

By organizing the documents according to feature standards, decision trees categorize the texts. In a decision tree, every node denotes an aspect that needs to be defined, and every branch denotes a possible value for the node to assume. Large trees may require assistance with overfitting; therefore, decision trees are typically constructed by expanding a tree to a respectable size and then trimming it back. Decision trees are easy to understand and require less pre-processing [19].

5) *Neural Networks:*

Deep Neural Networks (DNN):

DNN architectures follow a multilayer approach. It gains knowledge from numerous layer connections, in which each layer passes on connections to the subsequent layer in the concealed portion after receiving connections from the preceding layer. The output of DNN will be the multi-class classification of input text. Neural networks are employed in deep learning to boost computational effort and offer precise results [20].

Recurrent Neural Networks (RNN):

RNN is a type of neural network architecture applied for text classification where the output of the earlier step is provided as input to the next step.

Long Short-Term Memory (LSTM):

Short-Term Memory (STM) and Long-Term Memory (LTM) are concepts used in the context of Recurrent Neural Networks (RNNs), particularly with LSTM networks. Since LSTMs can capture long-range dependencies and retain information over time, they are an excellent sort of RNN for processing sequential input, including text. LSTMs are powerful in text classification tasks because they can handle varying-length sequences and effectively capture contextual information from the text. Numerous natural language processing (NLP) activities, such as sentiment analysis, entity identification, machine translation, and many more, have made extensive use of them.

Convolutional Neural Networks (CNN):

It is a multilayer artificial neural network that can detect and extract features from text data. CNNs were initially emerged for applications involving computer vision like image recognition, but they have also been adapted for text classification tasks with great success. In text classification, the goal is automatically assigning a label or category to a given text. Treating the text as a one-dimensional sequence of words or characters and applying one-dimensional convolutions to identify local trends or features is the key idea behind using CNN for classification. The convolutional filters slide over the input text, extracting features at different positions, and these features are then used for classification.

B. *Unsupervised document classification*

Unsupervised approaches, in contrast to supervised approaches, aim to categorise documents based on their differences rather than using a dataset for learning. Unsupervised techniques are more challenging than supervised techniques. Unsupervised techniques perform classification based on unlabelled data. It performs classification without human intervention. Unsupervised learning is preferable for the situations where data is unlabelled or it is difficult or expensive to label the dataset.

1) K-Means Clustering: It groups the data into group of K clusters based on similarity. The cluster with the closest centroid acquires the data. With this method, data is grouped into a nearby cluster.
2) Hierarchical Clustering: It generate a cluster tree to depict the hierarchical relationships within the data. This process involves classifying the data by traversing the clusters and assigning data at various levels, thereby unveiling the hierarchical structure of the dataset.

C. *Semi-supervised document classification*

Combining supervised and unsupervised techniques is known as semi-supervised learning. The semi-supervised approach can enhance both supervised and unsupervised document classification performance. It makes use of unlabelled data as well as a labelled training set. This is particularly useful when obtaining labelled data is expensive or time-consuming.

1) Self-training: It Train a classifier on the initially labelled data. It uses the trained classifier to predict labels for unlabelled data. It improves the performance of classifier with the help of pseudo-labelled data.
2) Co-training: It Trains multiple classifiers on different views of the data based on their features and representations. Here Each classifier predicts labels for unlabelled data.
3) Multi-view learning: It represents documents in multiple ways (e.g., bag-of-words, word embeddings) and use multiple classifiers to classify the data. It combines predictions from different classifiers and incorporate feedback from labelled instances to refine models.
4) Ensemble methods: It trains multiple classifiers independently by combining predictions from different classifiers. It Improves robustness and generalization by combining diverse models.

## V. CONCLUSION

The information in digital form is increasing rapidly; processing it for better utilization is essential. This paper reviews the various classifiers available for Indian Languages. As per the literature review it is found that the supervised techniques worked well on Indian Language text classification, but Indian content still needs to be explored in terms of text classifications. However, several text document classifiers are available for language classification, and more work is needed on Marathi Document classification.

It will be interesting to study the performance of various ML and DL document classifiers for classifying Marathi text documents.

## REFERENCES

[1]. “Design and Development of Marathi Text Classification System for Information Retrieval” [2017], Patil, R. P., Ph. D. thesis, at Kavayitri Bahinabai Chaudhari North Maharashtra University.

[2]. Murthy, K. N. (2003). Automatic categorization of Telugu news articles. Department of Computer and Information Sciences.

[3]. Ali, A. R., & Ijaz, M. (2009, December). Urdu text classification. In Proceedings of the 7th international conference on frontiers of information technology (pp. 1-7).

[4]. Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., & Palaniappan, B. (2009). Automatic classification of Tamil documents using vector space model and artificial neural network. Expert Systems with Applications, 36(8), 10914-10918.

[5]. R. Jayashree and M. K. Srikanta, "An analysis of sentence level text classification for the Kannada language," 2011 International Conference of Soft Computing and Pattern Recognition (SoCPaR), Dalian, China, 2011, pp. 147-151, doi: 10.1109/SoCPaR.2011.6089130.

[6]. Nidhi, et. al., "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.

[7]. Gupta, V., & Gupta, V. (2012). Algorithm for punjabi text classification. International Journal of Computer Applications, 37(11), 30-35.

[8]. Thaoroijam, K. (2014). A study on document classification using machine learning techniques. International Journal of Computer Science Issues (IJCSI), 11(2), 217.

[9]. Bam, S. and Shahi, T. (2014) Named Entity Recognition for Nepali Text Using Support Vector Machines. Intelligent Information Management, 6, 21-29. doi: 10.4236/iim.2014.62004.

[10]. Joulin, A., Grave, E., Bojanowski, P., &Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

[11]. Mandal, A. K., & Sen, R. (2014). Supervised learning methods for bangla web document categorization. arXiv preprint arXiv:1410.2045.

[12]. Puri, S., & Singh, S. P. (2019). An efficient hindi text classification model using svm. In Computing and Network Sustainability: Proceedings of IRSCNS 2018 (pp. 227-237). Springer Singapore.

[13]. Raju, G.V.S., Badugu, S., Akhila, V. (2022). Telugu Text Classification Using Supervised Machine Learning Algorithm. In: Bhateja, V., Satapathy, S.C., Travieso-Gonzalez, C.M., Adilakshmi, T. (eds) Smart Intelligent Computing and Applications, Volume 1. Smart Innovation, Systems and Technologies, vol 282. Springer, Singapore. https://doi.org/10.1007/978-981-16-9669-5_27

[14]. “Analysis of morbologically rich tamil language for document classification using machine laearing techniques”

[2021], RAJKUMAR N, Ph. D. thesis, at Annamalai University.

[15]. Deshmukh, R., Kiwelekar, A.W. (2022). Deep Convolutional Neural Network Approach for Classification of Poems. In: Kim, JH., Singh, M., Khan, J., Tiwary, U.S., Sur, M., Singh, D. (eds) Intelligent Human Computer Interaction. IHCI 2021. Lecture Notes in Computer Science, vol 13184. Springer,Cham. https://doi.org/10.1007/978-3-030-98404-5_7

[16]. Dhar, A., Mukherjee, H., Roy, K., Santosh, K. C., & Dash, N. S. (2023). Hybrid approach for text categorization: A case study with Bangla news article. Journal of Information Science, 49(3), 762-777.

[17]. Singh, K. N., Devi, S. D., Devi, H. M., & Mahanta, A. K. (2022). A novel approach for dimension reduction using word embedding: An enhanced text classification approach. International Journal of Information Management Data Insights, 2(1), 100061

[18]. Harshali B. Patil (2022),A Review of Document Classifier Systems Developed for Indian Languages, IJSRD - International Journal for Scientific Research & Development| Vol. 10, Issue 3, 2022 | ISSN (online): 2321-0613

[19]. “The Design and Development of Multi-Category Document Classification Al[1]gorithm using Text Mining in Gujarati Language” [2017] Rajnish M Rakholia, Ph. D. Thesis at Faculty of Technology, R K University, Rajkot.

[20]. “A framework for classification of medical data with deep learning based classification models" [2019], Lavanya Devi, G. Ph.D. Thesis at Andhra University.