ISSN:0975-9646

Text Categorization on Multiple Languages Based On Classification Technique

#[1]**Kapila Rani**
(M.tech Scholar)
[1]*Department of Computer Science and Engineering*
T.I.T, Bhiwani

*[2]**Satvika**
(Assistant Professor)
[2]*Department of Computer Science and Engineering*
T.I.T, Bhiwani

Abstract- In the Constitution of India, a provision is made for each of the Indian states to choose their own official language for communicating at the state level for official purpose [1]. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. Hence, the Classification of text documents based on languages is essential [2]. The objective of the work is the representation and categorization of Indian language text documents using text mining techniques. Several text mining techniques such as Support Vector Machine, KNN (K-Nearest Neighbor), Decision Tree, Self-Organizing Map(SOM), Genetic Algorithm [2].

Keywords- KNN Classification; Precision (p); Recall(r); F-measure; Tokens; Stop-words.

I. INTRODUCTION

Text categorization [1] can be briefly described as the automatization of the document organization process to a set of pre-defined categories. Automatic Text Classification is an important application and research topic for the identification of digital documents. A text classification system is used to index the documents for the information retrieval tasks and to the classification of memos, e-mails or web pages. Text Classification represents the high dimensionality of the feature space. The Text Classification is used to assign the category labels to the new documents at the training stage which are based on the knowledge gained in a classification system. In the training phase, a classification system is built using a learning method and a set of documents which are given, attached with class labels, machine learning communities.

Types of text categorization are as follows [20]:-

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:*
Given a document, the classifier search all the Categories to which the document belongs. This is stated as document-pivoted categorization. Like the classifier which searches all the documents that must be filed under a given category. This is stated as category-pivoted categorization.

C. Soft versus Hard Text Categorization:
Hard categorization is based on a binary decision made by automated categorization system on each document-category pair, while soft categorization means ranking the input documents or ordered the output categories is also called ranking categorization.

II. NEED FOR AUTOMATIC TEXT CLASSIFICATION

Due to the rapid increment in the collection of documents on internet, to classifying millions of text documents manually is an expensive and time consuming task. Therefore, automatic text classifiers are constructed using pre-classified sample documents whose accuracy and time efficiency is much better than manual text classification. In this paper text classification techniques that are used to classify the text documents into predefined classes.

Text classification has many applications in daily life because non-automated organization methods are not satisfactory anymore and there is a need to organize a enormous amount of data and documents automatically. Some examples of applications in daily life can be listed as follows:

a) Web page search engines use text categorization in a hierarchical manner. Documents are categorized according to their topics and chosen to be in a category listed hierarchically.
b) Text Categorization is used in text filtering. In this approach, a document is either forwarded or stopped in a flow to a user also called consumer- according to the consumer's profile. In this application, two categories are necessary.
c) Text Categorization is used in some Natural Language Processing (NLP) applications such as finding the meaning of a disambiguos word in a given text. Such words have more than one meaning and text categorization will be used in finding out, which meaning of the word is used in that sample text.

III. EXISTING WORK

Text categorization play an important role for several applications especially for organizing, classifying through few words representing large volumes of information. The Text Categorization goal is to label documents according to a predefined set of classes. Classification representation explains data relationships and predicts values for future observations. An extensive variety of techniques have been designed for text classification, such as decision tree

method Rule-based classifiers, Bayes classifiers, The nearest neighbor classifier, SVM classifier, neural network classifier and so on. In this paper K nearest neighbor Classifier are used for classification of the Indian language document the below figure is the existing work:-

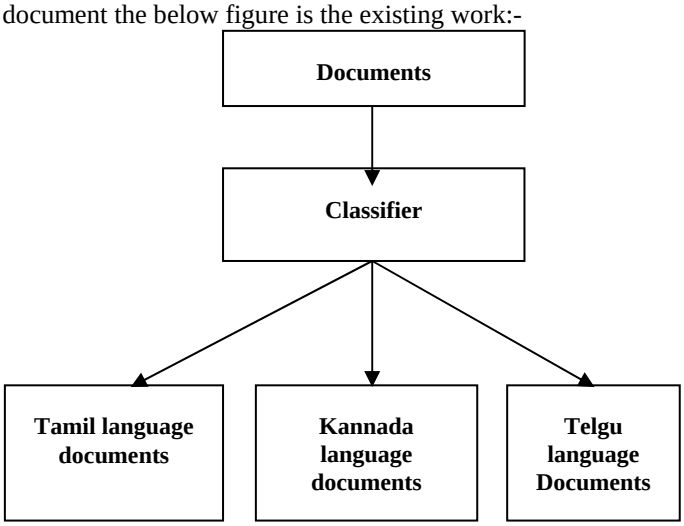


Figure3.1 Diagram of existing work

The use of removal algorithm k Nearest Neighbor (KNN) to south Indian languages such as Kannada, Tamil and Telugu text has been evaluated. In the existing algorithm, the corpus consists of 300 documents that belong to 3 categories. All the documents are preprocessed by removing stop words and light stem all the tokens. The documents are representing by means of the vector space model. For measuring the efficiency of organization algorithm, the traditional recall and precision measures is used. But the disadvantage is that the accuracy will be low.

Algorithm

a) Identify specific language files.
b) Associate a Language label with each of the files.
c) Build a Corpus C
d) Preprocess the Corpus C.
e) Generate VSM or expression article matrix using Binary Term Occurrence D (i, j) (where i is the document i and j is the jth term of document i.) (TF and TF-IDF are not used in the matrix because only the occurrence of the phrase in the DSL file is relevant for arrangement; the unique or rarity of the term is irrelevant in this approach)

IV. PROPOSED WORK

In existing work Telugu, Tamil and Kannada languages are used but in proposed English and Hindi language classifier are used. India is the home of different languages. Each state in India has its own official language. The objective of this work is to classify the documents based on language, using supervised learning algorithm. The main goal to use the more than one language is to increase the accuracy of classifier. Using the I-KNN algorithm the accuracy is much better as compared to existing algorithm and also calculates the f-measure, g-mean, precision, recall, accuracy of that algorithm. There are various steps used for preprocessing like documents, standardization, tokenization, remove stop words, stemming, vector space model.

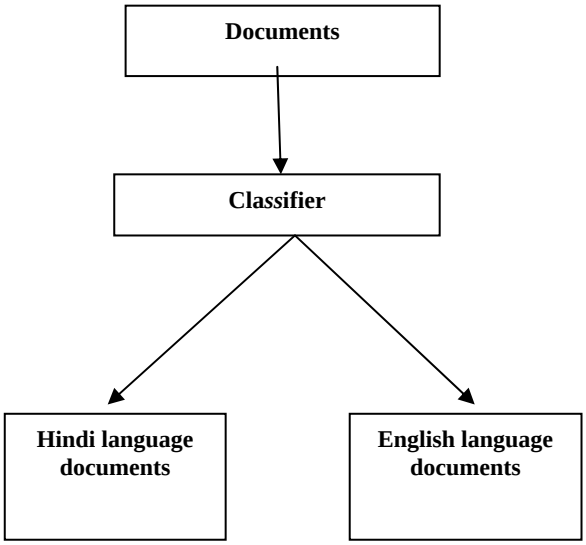


Figure4.1 proposed work diagram

TEXT REPRESENTATION

A. *Collecting Documents*

The work set aside for creating this corpus is the World Wide Web itself. The main problem with this approach to article collection is that the data may be alive of uncertain quality and require extensive cleansing before use.

B. *Document Standardization*

Once the credentials are collected, it is common to find them in a variety of different formats, depending on how the documents were generated. The documents should be processed with minor modifications to convert them to a standard format.

C. Tokenization

Given a character sequence and a specific document unit, tokenization is the task of chopping it up into pieces, called tokens, possibly by the side of the same time throwing away certain typescript, such as punctuation. The tokenization procedure is language dependent.

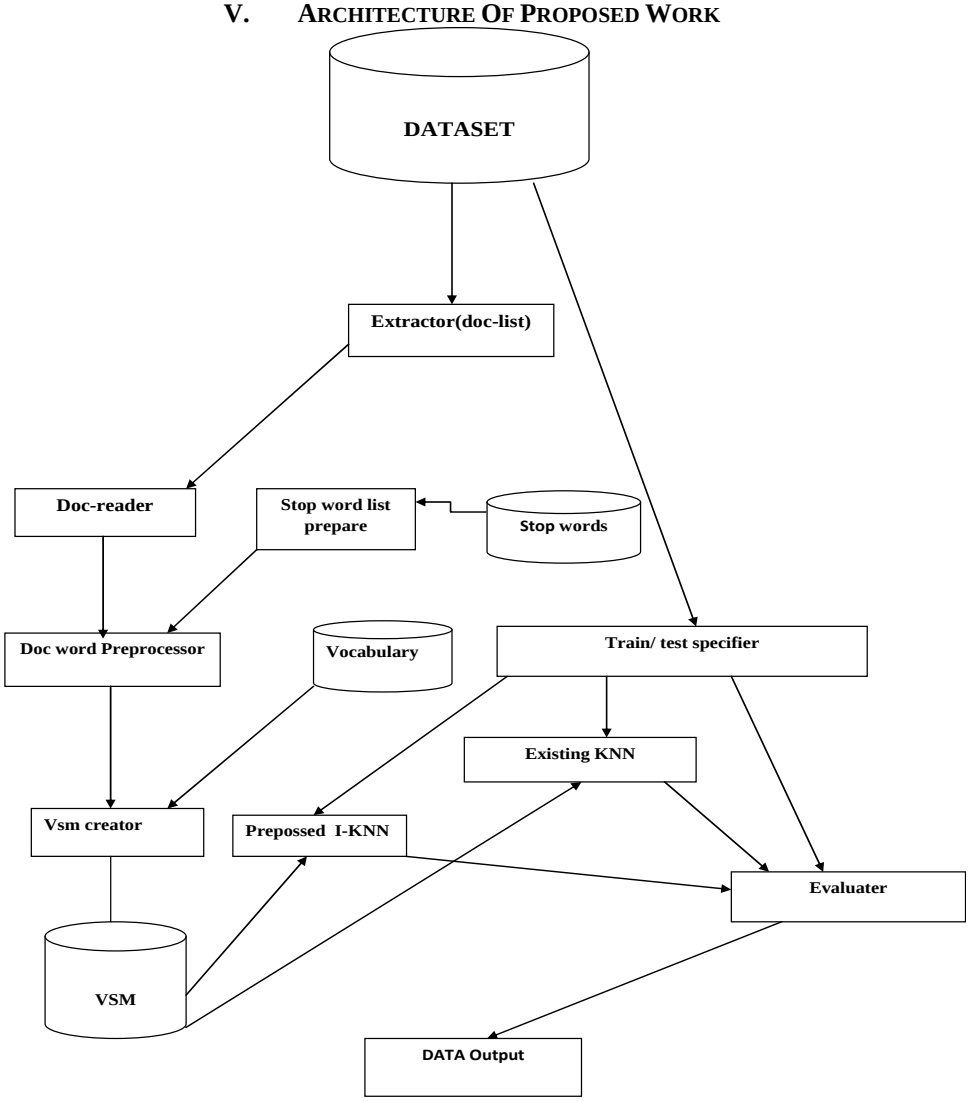
D. Dropping Common Terms

Stop-Words: Some enormously common words are not informative. These words are called stop-words. The strategy used for determining a stop list is to sort the terms by collection frequency (the total number of times every one term appears in the document collection), and then to take the nearly everyone frequent terms (stop language These words are discarded during indexing.

E. Lemmatization

Once tokens are fashioned, the next possible step is on the road to convert each of the tokens to a standard appearance a process usually referred to as stemming or lemmatization. The advantage of stemming is to reduce the number of distinct types in a text corpus and to augment the frequency of incidence of some human being types.

V. ARCHITECTURE OF PROPOSED WORK



Description of the proposed work:-

In proposed work two languages Hindi and English is used to build dataset. Dataset set have many files than extractor is used to rectify the file and then doc reader is used for reading the files. After that in which those words are repeated that's dataset have been created than it will create stop word list .After that doc-word reader and the stop word list prepare goes to the preprocessor. Using the technique of preprocessor the processing will be fast and that's words comes more than one time that have been removed through the preprocessing. Than the speed becomes fast. After the preprocessing vocabulary dataset has been created in which all words are come. Than that's goes to vsm (vector space model). Vector space model or term vector model is an algebraic replica for representing text documents as vectors. The principle behind the VSM(vector space model) is that a vector, with elements representing individual terms, may encode a document's meaning according to the relative weights of these term elements. After that train /test specified evaluator doing her work after that the existing and proposed results are occur.

VI. RESULTS

In which we will measure the accuracy, tpr/recall, precision, f-measure, g-mean.

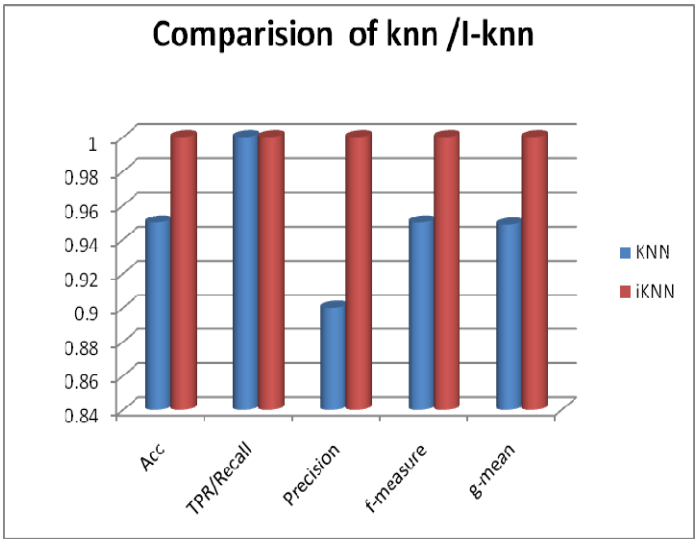


Fig: 6.1 Results of Hindi language classifier

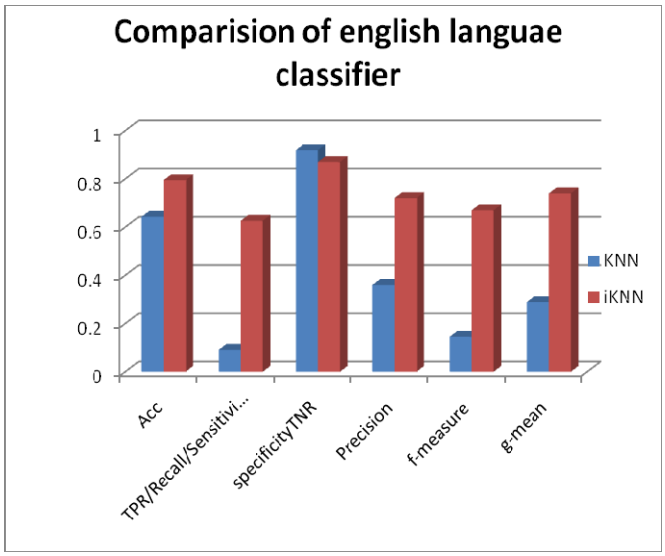


Fig: 6.2 Results of English language classifier

VII. CONCLUSION

The classification problem is one of the most fundamental problems in the text categorization. There are many techniques for Text Categorization such as decision tree, Bayes method, nearest neighbor classifiers, and SVM classifiers. But this paper discusses only k nearest neighbor technique .The overall conclusion is that in existing work Telugu, Tamil and Kannada languages was used. But in proposed work the Hindi and English languages test by the java tool. According to the proposed work the accuracy is much better from the existing work. They were used the KNN classifier but this work uses the I-KNN classifier to improve the accuracy much better. To improve the accuracy this paper uses the various feature selection methods like Tf(term frequency), Idf(Inverse Document Frequency) and measuring the categorization effectiveness f-measure and also calculate actual result and draw the graph., recall, precision, g mean, accuracy, TPR, TNR

VIII. FUTURE WORK

In future there will be various techniques can be applied on the text categorization. In future we will use the various more algorithms to check their accuracy and become larger dataset and apply on more languages. But this paper uses the text categorization on Hindi and English language using java and tests their accuracy using KNN and I-KNN classifiers. In future Naïve Bayes, Back propagation Neural Network, Latent semantic indexing using CNN, Support Vector Machine, KNN (K- Nearest Neighbor), Decision Tree, Self-Organizing Map (SOM) and Genetic Algorithm is used to find the better accuracy [2].

References

[1] M Narayana Swamy"Indian Language Text Representation and Categorization using Supervised Learning Algorithm'', International Journal of Data Mining Techniques and Application, December 2013.

[2] Text Categorization on Multiple Languages Based on Classification Technique, International Journal for Scientific Research & Development| Vol. 3, Issue 03, 2015

[3] B. Mahalakshmi "An Overview of Categorization techniques", International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.5, Sep.-Oct. 2012.

[4] M.Aly "Survey on Multiclass Classification Methods," Neural Networks, 2005.

[5] M. E. Ruiz and P. Srinivasan, "Automatic Text Categorization Using Neural Networks," School of library and Information Science, Vol.8.

[6] W.and B. Yu, "Text categorization based on combination of modified back propagation neural network and latent semantic analysis," Neural Compute & Applic, Springer Link, Vol. 18, No.8, pp.875–881, 2009.

[7] Text Mining and Scholarly Publishing, Jonathan Clark, Publishing Research Consortium 2013, PRCTextMining and Scholarly Publish in Feb2013 Salton G and McGill M 1983 Introduction to Modern Information Retrieval .McGraw-Hill, NewYork.

[8] Representation and Classification of Text Documents: A Brief Review B S Harish, D S Guru, S Manjunath" IJCA Special Issue on "Recent Trends in Image Processing and Pattern. Recognition"RTIPPR, 2010."

[9] D.D. Lewis and M. Ringuette, " A comparison of two learning algorithms for text categorization, Symposium on Document Analysis and Information Retrieval,"1994.

[10] Mr. B. R Patel, Mr. K.K Rana," A Survey on Decision Tree Algorithm For Classification,"Volume 2, Issue 1, ISSN: 2321-9939, 2014.

[11] C. Hua Li and S. Choel Park,"Text Categorization Based on Artificial Neural Networks, "Neural Information Processing, Springer, Vol.4234, pp.302-311, 2006.

[12] R.Freeman, Hujun Yin and N. M.Allinson,"Self-Organizing Maps for Tree View Based Hierarchical Document Clustering,"IEEEXplore,pp.1906-1911,2002.

[13] Y. Yu, Pilian He, Y. Bai and Z. Yang, "A Document Clustering Method Based on One-Dimensional SOM,"Seventh IEEE/ACIS International Conference on Computer and Information Science, pp.295-300, 2008.

[14] Ganatra, Y P Kosta, G. Panchal and C. Gajjar, "Initial Classification Through Back Propagation In a Neural Network Following Optimization Through GA to Evaluate the Fitness of an Algorithm," International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[15] E. Youn and M. K. Jeong, "Class dependent feature scaling method using naive Bayes classifier for text datamining," Pattern Recognition Letters, vol. 30, pp. 477-485, 2009.

[16] S. S. R. Mengle and N. Goharian, "Ambiguity Measure Feature-Selection Algorithm," Journal of the American Society for Information Science and Technology, vol. 60, pp. 1037-1050, May. 2009.

[17] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra category for text categorization," Information Processing and Management, vol. 48, pp. 741-754, Jul. 2012.

[18] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in Machine Learning: ECML-98. vol. 1398, C. Nédellec and C. Rouveirol, Eds., ed: Springer Berlin /Heidelberg, 1998. .

[19] Pooja Bolaj et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (2), 2016.