I J C T A, 9(3), 2016, pp. 1473-1481
© International Science Press

Indian Language Text Documents Categorization and Keyword Extraction

Hanumanthappa* and Narayana Swamy M.**

Abstract: Now a day's lot of Indian language content is being generated in digital form. Managing a vast amount of documents in digital forms is very important in text mining applications. Indian language Text Mining has become an important research area. Text Mining means extracting knowledge from text documents Text categorization and summarization are the two important techniques to extract the knowledge.

Text Categorization is the task of automatically sorting a set of documents into categories based on language. Document Summarization is an emerging technique for understanding the main purpose of the content of the documents. This paper presents a model that uses text categorization based on language and text summarization by extracting keywords.

Keywords: *Text Mining, Text Categorization, TFIDF*

1. INTRODUCTION

India is the home of different languages, due to its cultural and geographical diversity. 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. In India, the growth in consumption of Indian language content started because of growth of electronic devices and technology. The availability of constantly increasing amount of textual data of various Indian regional languages in electronic form has accelerated. But not much work has been done in Indian languages text processing. So there is a huge gap from the stored data to the knowledge that could be constructed from the data. This research is concerned with the study and analyzes the text mining for Indian regional languages

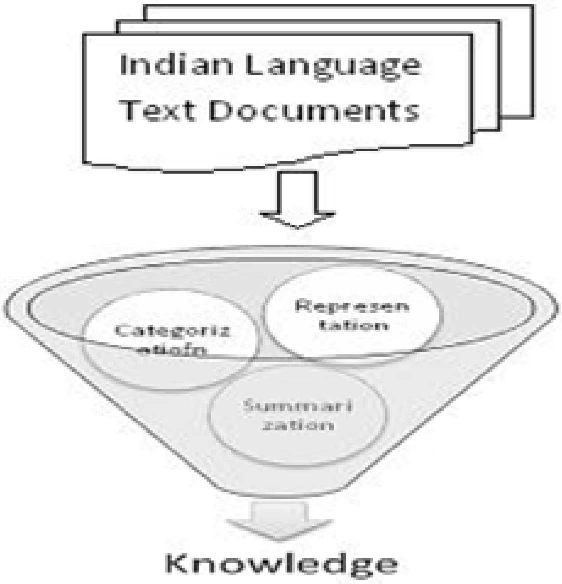
Now a day's in India internet services is provided not only in English but also available in regional language. Therefore lot of Indian language content is being generated in digital form.Huge number of available documents in digital media makes it difficult to obtain the necessary knowledge related to the needs of a user. So with exponential increase in the information in Indian languages on the web, automatic information processing and retrieval become an urgent need. This motivated us to work on Text mining for Indian languages [1]. To utilize the available text data, in this paper we have proposed a techniques to represent the Indian language and a model to extract the knowledge by using text mining techniques such as text categorization and keyword extraction.

2. LITERATURE REVIEW

From the literature survey noticed that, not much work has been done to capture various aspect of knowledge from Indian language text. Here an attempt is made to summarize the research work on Indian languages text processing. The topic tracking for Punjabi language has been experimented with two approaches. NER (Name Entity Recognition) based approach and keyword extraction approaches have been implemented [2]. Part of speech (POS) tagging plays a vital role in natural language processing. This paper presents a reasonably

* Professor Department of Computer Applications Bangalore University Bangalore, *Email: Hanu6572@hotmail.com*

** Research Scholar Bharatiyar University Coimbatore Tamil Nadu, *Email: Narayan1973.mns@gmail.com*



accurate POS tagger for Kannada language [3]. Ontology Based Classification and Hybrid Approach (which is the combination of Naïve Bayes and Ontology Based Classification) are used for Punjabi Text Classification [4]. Single-document opinion summarization for Bengali, in this work the proposed technique is the topic based document-level theme relational graphical representation.[5]. For the first time resources have been developed for Punjabi and these can be beneficial for developing other Natural language processing applications for Punjabi [6]. Document Summarization In Kannada Using Keyword Extraction. The summarizer can be used as a tool in various organizations such as Kannada Development Authority, Kannada SahityaParishathetc [7]. To extract relevant data from Oriya language using Classification algorithms C 5.0 [8]. Classified Telugu documents using Naive Bayes classifier. The base system on which a variety of further explorations can be carried out, both from the linguistic point of view and statistical point of view[9].

3. MOTIVATION

India is one of the multilingual nations in the world today. India's growing focus on Internet services being provided in regional languages. The first step in this direction was the launch of TDIL (Technology Development for Indian Languages) Programme in 1991 by Ministry of Information Technology to develop information processing tools to facilitate human machine interaction in Indian Languages.

[10] States that the internet users in India could increase by 24% if local language content is provided on the internet. Amongst current active internet users, local languageusage

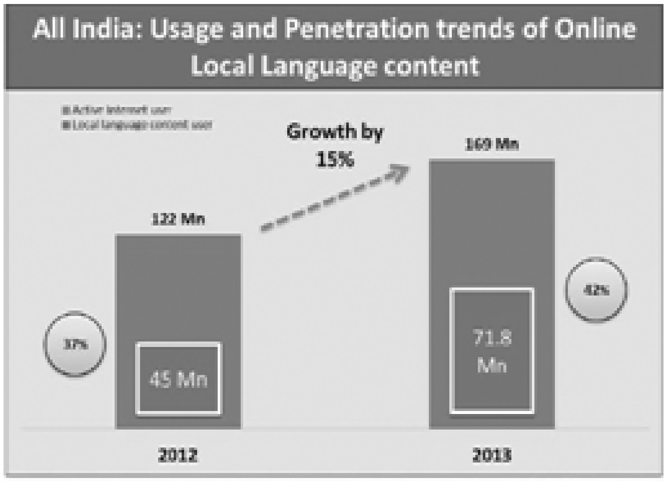


Figure 1

Base: All India Active Internet Users (Urban), December 2013
All India Active Internet Users (Urban), June 2012

Penetration is around 42%t. Huge number of available documents in digital media makes it difficult to obtain the necessary knowledge related to the needs of a user. So with exponential increase in the information in Indian languages on the web, automatic information processing and retrieval become an urgent need. This motivated us to work on Text mining for Indian languages.

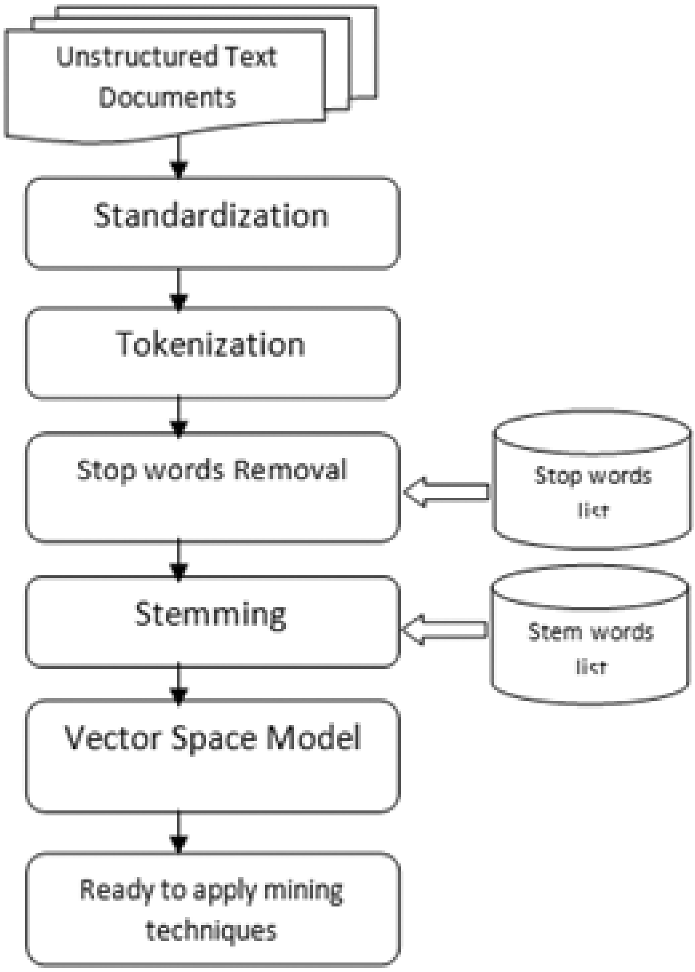
4. OBJECTIVE

- 4.1 Data preprocessing
- 4.2 Document categorization
  - o Based on the language
  - o Based on the domain
- 4.3 Keywords extraction

4.1. Data Preprocessing

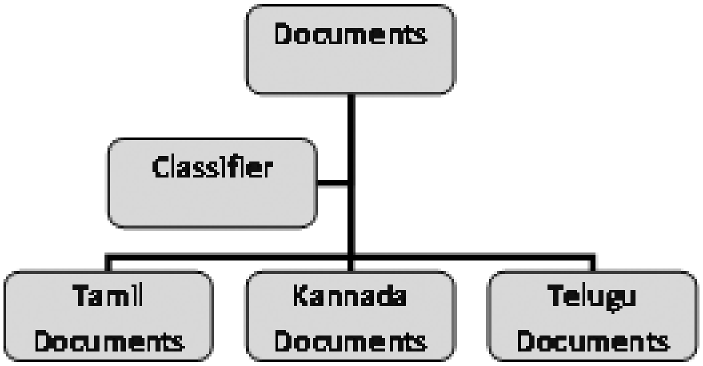
Data Preprocessing means converting unstructured data into structured data. Given a textual source containing different types of documents (different formats, language formatting) the first action that should text preprocessing.[11]The preprocessing steps are as follows

The last step of preprocessing is representing documents using vector space model. Here the document is represented in the form of matrix by using the term frequency, Document frequency, Inverse Document frequency and TI-IDF. After preprocessing data mining algorithms can be applied.



4.2. Document Categorization

Categorization is the process of dividing the data into number of groups on the basis of a training set of data. Categorization is an instance of supervised learning, i.e. learning where a training set of correctly identifiedobservations is available.The task of categorization can be done by using several methods using different types of classifiers. Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems. Determining a suitable classifier for a given problem is however still more an art than a science.



4.2.1. Document Categorization Based on the language

Algorithm

Step 1: Identify specific language files.

Step 2: Associate a Language label with each of the files.

Step 3: Build a Corpus C

Step 4: Preprocess the Corpus C.

Step 5: Apply a Stemming algorithm to reduce all the words to their root form.

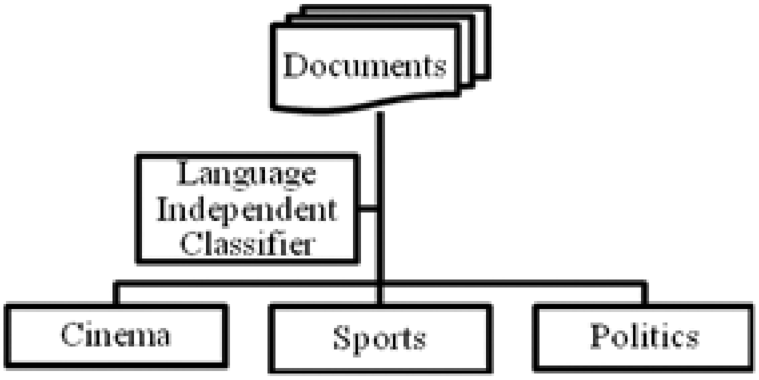
Step 6: Generate VSM or a Term Document matrix using Binary Term Occurrence D(i, j)

(where i is the document i and j is the jth term of document i.)

Step 7: Train the Classifier (kNN, j48 and NB) using C as training examples.

Document Categorization Based on the Domain

The digital data is available in all the Indian languages. It would be too expensive to build an individual system for each language separately. The objective of this work is to design a language independent classifier to categories the documents based on domain. In this work we have used supervised learning algorithm like



Decision trees and K-nearest neighbor. The corpus is created for three domain cinema, sports and politics from WWW.

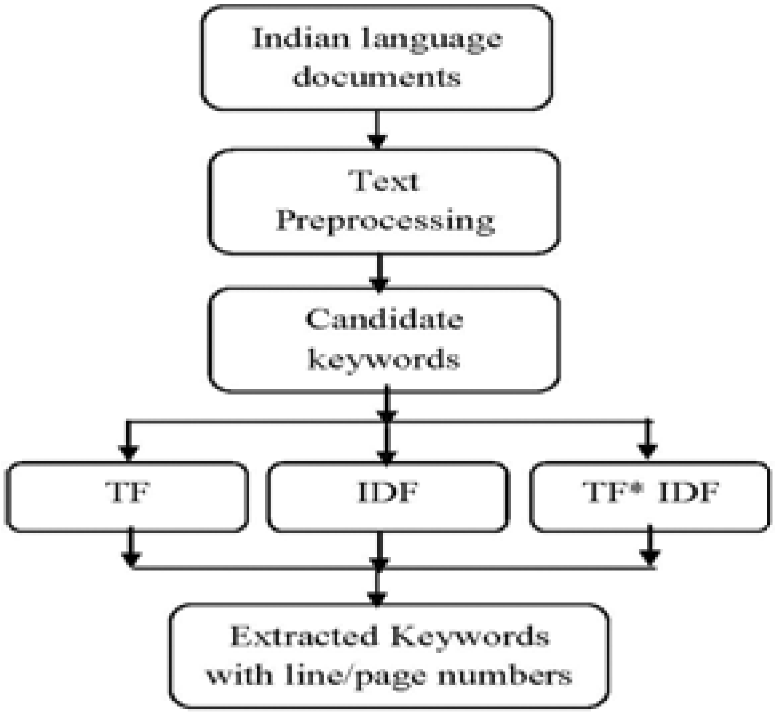
4.3. Keywords extraction

Keywords are widely used as a brief summary and index of documents. Keyword extraction is the task selecting a small set of words from the document that can describe the meaning of the document. In this works, the term frequency – inverse document frequency alalso called TF*IDF, is a used to evaluate howimportant is a word in a document.

TF – The weight of a term that occurs in a document is simply proportional to the term frequency.

IDF – The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs**.**

TF*IDF – The product of two statistics, term frequency and inverse document frequency. The **inverse document frequency** is a measure of how much information the wordprovides, that is, whether the term is common or rare across all documents.


Indian language documents
Text Preprocessing
Candidate keywords
TF
IDF
TF* IDF
Extracted Keywords with line/page numbers

5. RESULT AND PERFORMANCE

5.1. Document Categorization

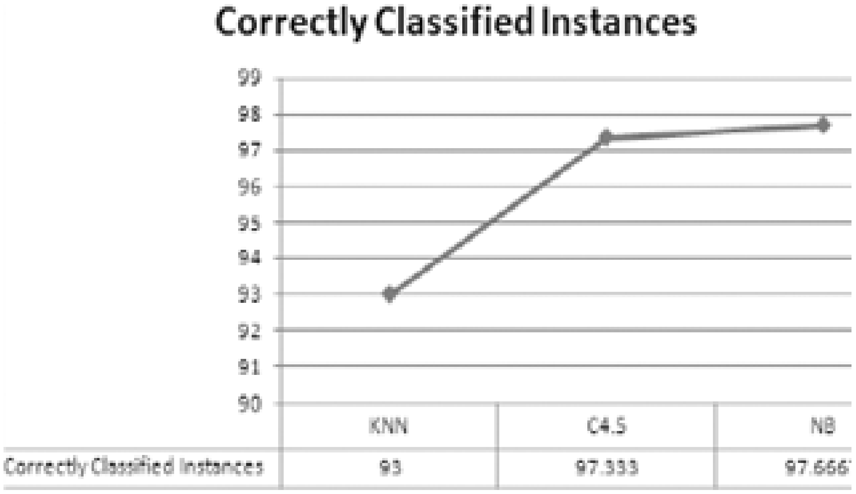
The prediction of the **Document categorization models**are tabulated in the form of confusion matrix as shown in table 2. The accuracy of the models are shown in the fig 3 the kNN classifier; predicted 87 Kannada document correctly and 13 incorrectly, predicted 96 Tamil document correctly and 4incorrectly, predicted 96 Telugu document correctly and 4 incorrectly.

So the correctly classified documents are 93%. Same way the C4.5 classifiers gives 97.333% accuracy and NB classifier gives 97.666% accuracy. From figure 5, it can be observed that NB Classifier gives better result in comparison to kNN and C4.5(J48) Classifier

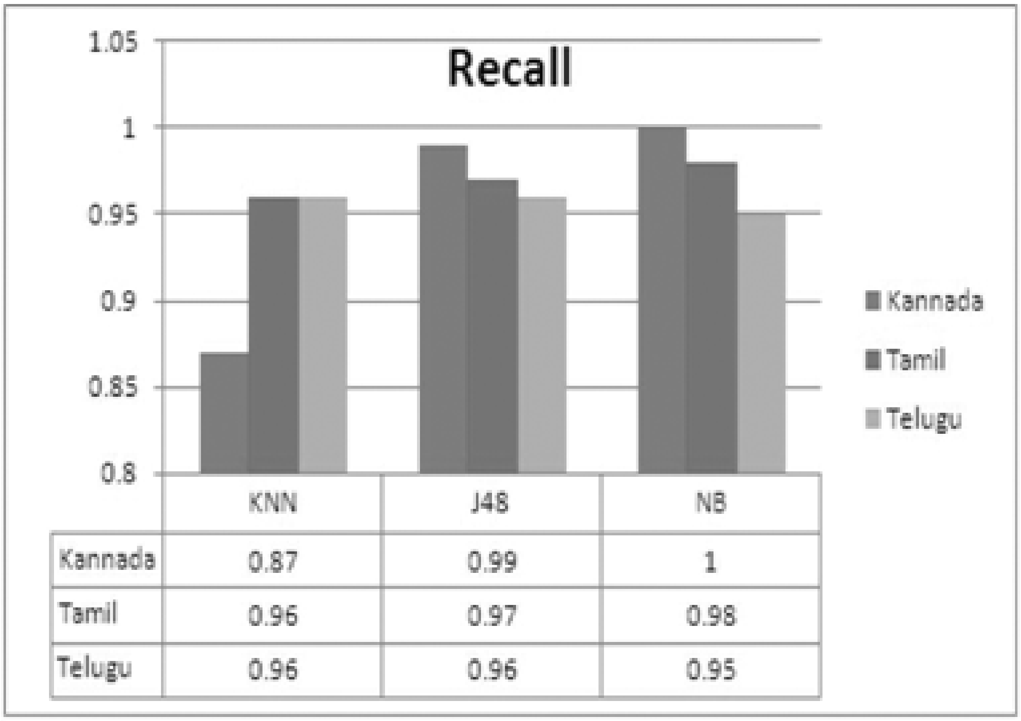
The effectiveness of a text classifier can be evaluated in terms of its precision (p), recall (r) and F-measure.

Table 1

Confusion Matrix								
	KNN Classifier			*J48 Classifier*			*NB Classifier*	
Kannada	*Tamil*	*Telugu*	*Kannada*	*Tamil*	*Telugu*	*Kannada*	*Tamil*	*Telugu*
87	2	11	99	1	0	100	0	0
2	96	2	1	97	2	2	98	0
4	0	96	4	0	96	5	0	95



	KNN	C4.5	NB
Correctly Classified Instances	93	97.333	97.666



	KNN	J48	NB
Kannada	0.87	0.99	1
Tamil	0.96	0.97	0.98
Telugu	0.96	0.96	0.95

5.2. Document Categorization Based on the Domain

We have considered two cases. In the first case only few documents are considered, but in case 2 quite a large amount of documents are considered for categorization as shown in the table 3.For classification problems it is natural to measure a classifier's performance in terms of the error rate. The error rate is calculated by using confusion matrix. The prediction of the classify models for two case are tabulated in the form of confusion matrix.

Table 2: KNN Classifier

Case 1

	Predicted Class			
	Cinema	Sports	Politics	Total
Actual Class	0	0	5	5
	0	0	5	5
	1	0	4	5
Total	1	0	14	

Table 3: J48 Classifier

	Predicted Class			
	Cinema	Sports	Politics	Total
Actual Class	2	0	3	5
	0	3	2	5
	0	1	4	5
Total	2	4	9	

Case 1: kNN Classifier accuracy = (0+0+4) / (0+0+5+0+0+5+1+0+4) = 26.66% J48 Classifier accuracy = (2+3+4) / (2+0+3+0+3+2+0+1+4) = 60.00%

Table 2: KNN Classifier

Case 2

	Predicted Class		
	Cinema	Sports	Total
Actual Class	38	10	48
	2	46	48
Total	40	56	

Table 3: KNN Classifier

	Predicted Class		
	Cinema	Sports	Total
Actual Class	48	0	48
	28	20	48
Total	76	20	

Case 2: kNN Classifier accuracy = (38+46) / (38+10+2+46) = 87.5%J48 Classifier accuracy = (48+20) / (48+0+28+20) = 70.83%

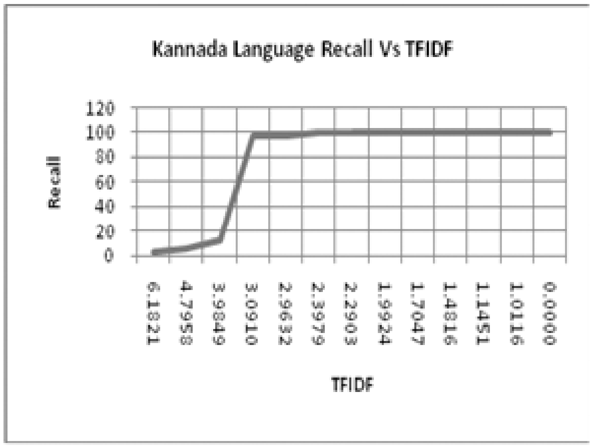
5.3. Keyword extraction

Three languages documents are selected to extract the keywords. Then the term frequency – inverse document frequency also called TF*IDF, is a used to evaluate how important is a word in a document.

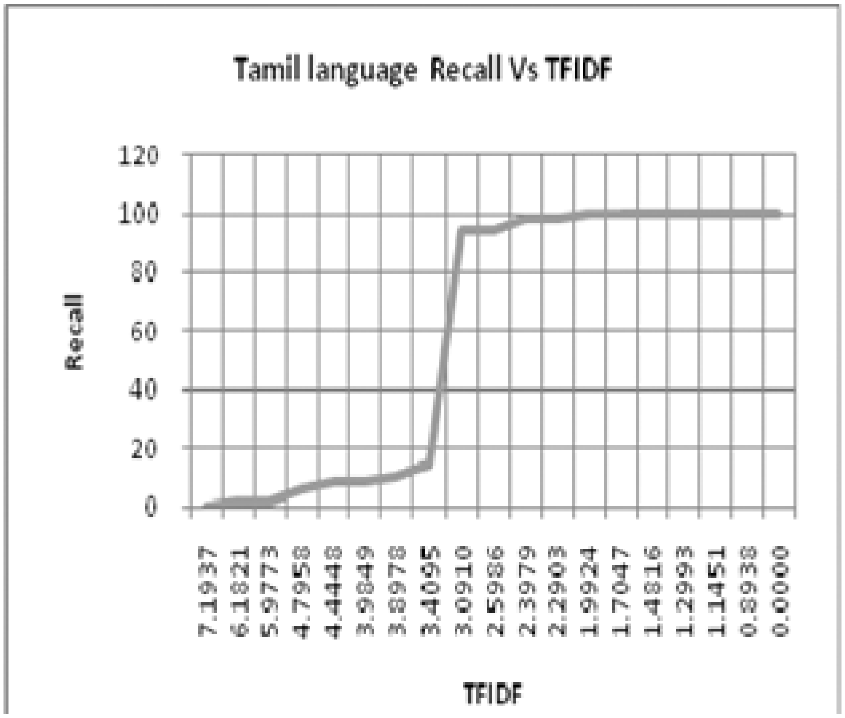
For keyword extraction Precision and recall are evaluation metric. Precision is the proportion of returned keywords that are targets, while recall is the proportion of target keywords returned. This Keyword extraction method is language independent, but the selection of threshold value will have a great impact on the result. From the evaluation of the three language, we can see that a good threshold value would be somewhere between 2.5 and 3.5.

In the case of Kannada text, when TFIDF is 3.0910 the recall is 100%.

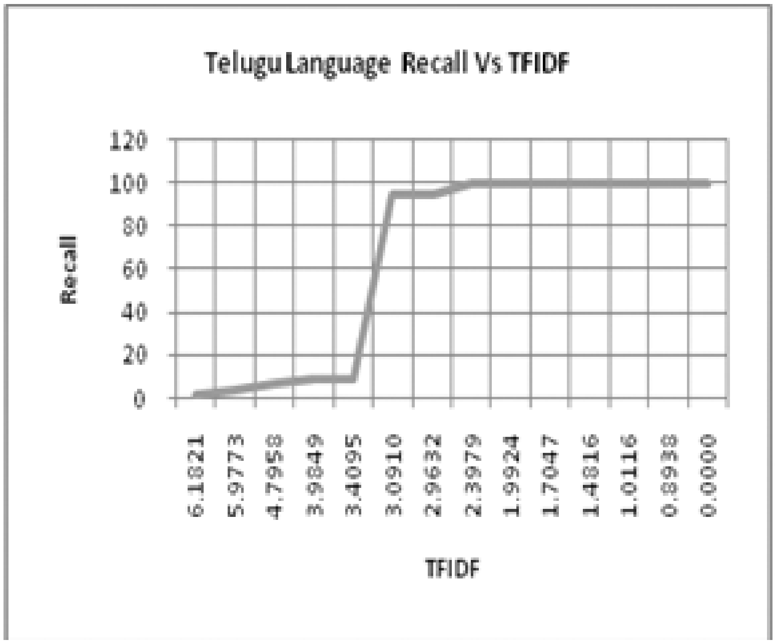
Therefore 3.0910 is used as a TFIDF threshold value to extract the keywords. So by considering those words whose TFIDF value is grater then 3.0910 as keywords.



In the case of Tamil text, when TFIDF is 2.3979 the recall is 100%. Therefore 2.3979 is used as a TFIDF threshold value to extract the keywords. So by considering those words whose TFIDF value is grater then 2.3979 as keywords.



In the case of Telugu text, when TFIDF is 3.4095 the recall is 100%. Therefore 3.4095is used as a TFIDF threshold value to extract the keywords. So by considering those words whose TFIDF value is grater then 3.4095 as keywords



6. CONCLUSIONS

An attempt is made to categorize the Indian language text documents based on language and domain. Aa common algorithm is designed to extract keywords from all the Indian languages. The biggest challenge is to converting unstructured data into structured data. Vector space model (VSM) is used to represent the text in the form of matrix. For document categorization based on language and domain, the algorithms such as, K nearest neighbor(KNN) approach, Decision tree and naive bayes classifier are used. The results are satisfactory. There is a scope to improve the accuracy.

The third phase of work is keyword extraction. In this work, TF*IDF is a used to evaluate how important is a word in a document. The TF*IDF is used as a threshold to select the important keyword. In the case of Kannada text, when TFIDF is 3.0910 the recall is 100%. Therefore 3.0910 is used as a TFIDF threshold value to extract the keywords. In the case of Tamil text, when TFIDF is 2.3979 the recall is 100%. Therefore 2.3979 is used as a TFIDF threshold value to extract the keywords.In the case of Telugu text, when TFIDF is 3.4095 the recall is 100%. Therefore 3.4095 is used as a TFIDF threshold value to extract the keywords.

References

[1] "Indian Language Text Mining" published in the international *Journal of Software Engineering and Simulation Volume 2 ~ Issue 10 (2015) pp: 01-04 ISSN (Online): 2321-3795 ISSN (Print): 2321-380. On May 2015.* Published with open access at www.questjournals.org

[2] "Topic Tracking For Punjabi Language", KamaldeepKaur and Vishal Gupta, Computer Science & Engineering: An International Journal (CSEIJ), Vol. 1, No. 3, August 2011 DOI : 10.5121/cseij. 2011.1304 37

[3] "POS Tagger for Kannada Sentence Translation", Mallamma V Reddy and Dr. M. Hanumanthappa, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 1, May-June 2012 ISSN 2278-6856

[4] "Algorithm for Punjabi Text Classification", Nidhi and Vishal Gupta, International Journal of Computer Applications (0975–8887) Volume 37– No. 11, January 2012.

[5] "Topic-Based Bengali Opinion Summarization", Amitava Das and SivajiBandyopadhyay, Coling 2010: Poster Volume, pages 232–240, Beijing, August 2010.

[6] "Automatic Punjabi Text Extractive Summarization System", Vishal Gupta and Gurpreet Singh Leha, Proceedings of COLING 2012: Demonstration Papers, pages 191–198, COLING 2012, Mumbai, December 2012.

[7] "Document Summarization In Kannada Using Keyword Extraction" Jayashree.R, SrikantaMurthy.K and Sunny.K, Computer Science & Information Technology (CS & IT)

[8] "Oriya Language Text Mining Using C5.0" Algorithm SohagSundar Nanda, Soumya Mishra, SanghamitraMohantyISSN : 0975-9646

[9] "*Automatic Categorization of Telugu News Articles*", KaviNarayanaMurthy, Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, 500 046http://202.41.85.68/knm-publications/il_text_cat.pdf.

[10] *Local Language Study 2013*, published jointly by Internet and Mobile Association of India (IAMAI) and IMRB International

[11] "*Indian Language Text Representation and Categorization using Supervised Learning Algorithm"* in the Proceedings of the International Conference on Intelligent Computing Applications (ICICA-14), organized by Bharathiar University, Coimbatore, Tamilnadu, during 6th and 7th March 2014. This attached with IEEE Affiliation published by IEEE -CPS and the corresponding ISBN 978-1-4799-3966-4. The copyright of the paper is owned by IEEE.