www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 3 March 2017, Page No. 20774-20779
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i3.67

Text Classification Using Centroid Technique

Shalini[1], Ubeeka Jain[2]
[1] Research Scholar
[2]Faculty of Computer science and Information Technology, I.K. Gujral Punjab Technical University, Rayat Institute of Engineering and Information Technology, India
[1]shallu.jassi24@gmail.com , [2]ubeekajain@gmail.com

Abstract- **During these days, till now there is no classifier used for classification of Punjabi documents. Here we have some Punjabi news article facts, which we have necessity to examine with the use of algorithms. Punjabi is an Indo Aryan language spoken in west Punjab (Pakistan) and East Punjab (India). So just a little work has been done in Text Classification for Punjabi language. In this work, assigns to the centroid technique for text News classification to construct structure for Punjabi Language. Huge News is accustoming for training and testing persistence of the classifiers. Language detailed pre-processing technique are beneficial for raw data to originate an identical or reduced-feature lexicon. Punjabi language is a morphological amusing language which makes those responsibilities complicated and hard to do. Statistical physical characteristics of bulk and lexicon are dignified which demonstrate satisfactory and appropriate results and consequences of text pre-processing of each distinct portion. We are residue able to get satisfactory and sufficient out comes or exertion using centroid Classifier. This research paper to reveal the centroid text News classification system to fix firmly for Punjabi Language. News quantity is accustomed for training and analysis purpose of the classifiers. Punjabi language is morphological rich language which originates those jobs and trustworthy complex. Statistical features of body and lexicon are measured which show acceptable results of text pre-processing section. We are able to get appropriate/competent out comes using centroid Classifier Algorithm**

.Keyword- Text classification, Punjabi language, Centroid Classifier.

I. INTRODUCTION

Textual classification is a process to arrange an assemblage of documents automatically into definite notations from a predetermined set. Nowadays, almost all of the efficacious items are in digital form, and taking attention of such data become hard. Consequently, text is to arrange the document into predefined classes, this lead to grow access rate and efficiency of the search engine. Manual text distinction is a valuable and time-consuming method, as it is come to be hard to arrange hundreds of thousands of documents manually. Therefore, automatic text classifier is built the use of categorized files and its accuracy is a lot higher than manual text type and it is much less time consuming too. Text classification is conceal in number of petitions such as report corporation, inquiry of exciting data, text filtering. The primary intention of text mining is to admit customers to deduce facts from textual assets and bargain with the manipulations in the same manner such as, to revive, classification is in supervised, unsupervised and semi supervised manner and summarization of natural Language Processing (NLP), assertion Mining, and gadget gaining cognition of strategies paintings together to robotically arrange and find out patterns from the exclude forms of the documents. For example, as an instance could be to mechanically label every incoming information tale within topic such as “sports activities”, “politics”, or “artwork”. A statistics mining classification assignment to give away off evolved within a training set E = (e1….. En) of documents which might be previously labelled within a category D1, D2 (e.g. recreation, politics). [1] The challenge is then to decide a class structure that's able to assign the best elegance to a new report e of the area. Textual content classification has one and one flavours as particular label and multi-label. Particular label is to be connected with single class and multi label document can to be connected with multiple classes.

Text Classification process is one and one -step procedures:

I. Training Phase: the accumulation of documents in this process is called training set. Every document in the training set appertain to specific class depends on their contents, called Labelled Documents. Training Phase deduct text classifier to categorize the unlabeled documents.

II. Testing Phase (also called Classification Phase): In this section, unlabeled documents are categorized accustom of labelled files. And to estimate the accuracy of the classifier, known class of the unlabeled document is similitude with classification result. [2]

There are two and one phase for processing:--

a. Pre-processing phase.
b. Feature extraction phase.
c. Processing phase.

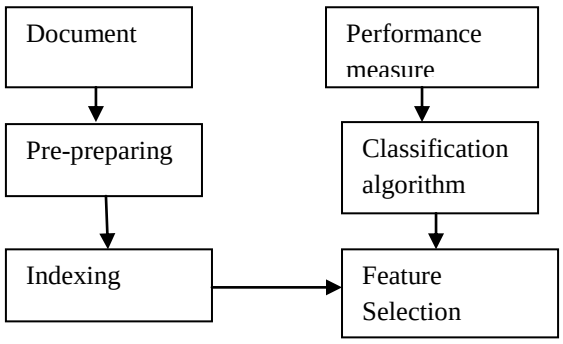

Fig 1. Text Classification Process

a. Pre-processing phase

The beginning pace of pre-preparing which is accustoming to attentive the set of records into clear word type. The set of record adapt for coming step in text classification is conducted by a large numbers of exhibiting peculiar quantities. Ordinarily the paces taken are:

- **Tokenization:** A inscription is behaving as a string, and then split into a catalogue of tokens.
- **Deleting stop words:** Stop words for instance "the", "a", "and", and so on are often happening, so the inconsequential letters should be expelled.

- **Stemming word:** Applying the stemming calculation that translates distinctive word structure into comparable standard structure. This progression is the way toward conflating tokens to their root structure, e.g. correction to correct, computing to compute.

b. Feature Extraction

Behind pre-preparing and indexing the vital advancement of text classification, is feature selection to build vector space, which enhances the quality of adaptation, the quality of being effective and precision of a content classifier. The fundamental conception of Feature Selection (FS) is to select subset of essential component from the initial archives. FS is to be executed by restraint the words with most astounding score as to suggest by foreordained measure of the impressiveness of the word. For this reason, text classification is greater egress is the honourable dimensionality of the feature space. Numerous feature evaluation metrics have been conspicuous in comparison with which are information gain (IG), term frequency, Chi-square, reversionary cross entropy, Odds Ratio, the weight of testimony of text, reciprocal information, Gini index.. For textual classification a major issue is the intense dimensionality of the feature space. Almost each text domain has more number of component, a large portion out of these elements are denial significant and yet gainful excess text classification specific work, and uniform some noise features may strongly decrease the characterization precision. Hence FS is to be regularly accustomed in text classification to decrease the magnitude of appearance space and enhance the productivity and exactness of classifiers. In stemming, distinctive types of the same word are united into a solitary word. For instance, particular, plural and distinctive tenses are solidified into a solitary word. We take note of that these strategies are not particular to the instance of the arrangement issue, and are regularly utilized as a part of an assortment of unsupervised applications, for example, bunching and indexing. On account of the order issue, it makes sense to supervise the feature selection process with the utilization of the class names. This sort of choice procedure guarantees that those elements which are exceptionally skewed towards the nearness of a specific class name are picked for the learning procedure.

II. LITERATURE REVIEW

In 2003, Kavi Narayana Murthy, "Automatic Categorization of Telugu News Articles", Department of Computer and Information Sciences, University of Hyderabad, Hyderabad,
In this there are automatically sorting a collection of document into categories that are from predefined set.

In 2008, Yutaka Sasaki NaCTeMSchool of Computer Science used Automatic Text Classification. Text Classification is the process to sort out documents into predefined classes
Text Classification is also called Text Categorization, Document Classification and Document Categorization. [3]

In 2009, Ali, Abbas Raza and IjasMaliha.Urdu Text Classification. Here we have some Urdu news article examples which we need to classify by making use of algorithms.

In 2009, Chen Jingnian, Huang Houkuan, Tian Shengfeng and Quyouli. Feature selection for text classification with Naive Bayes. In: Expert Systems with Applications. In this text categorization play important role in machine learning, text mining and information retrieval. It has been successful in handling large types of word. In this they have discussed various types of techniques and strategies for text categorization of Indian dialect. [4]

Rajan, K. Ramalingam, V. Ganeshen, M. Palanivel, S. And Palaniappan, B. (2009). Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network. In: Expert Systems with Applications. In this there are automatically sorting a collection of document into categories that are from predefined set.

In 2009,V.Ganesan, M., Palanivel, S., Palaniappan, Ramalingam, Rajan , "Automatic Classification of Tamil documents using Vector Space Model and Artificial Neural Network" Expert Systems with Applications.

In 2010, B S Harish, D S Guru and S Manjunath Representation and Classification of Text Documents. This paper gives a short summary to the many text representation schemes and classifiers used in the branch of text mining. The existing techniques are compared and contrasted depending on many standards namely criteria used for classification, algorithms adopted and classification time complexities. [5]

III. RELATED WORK

Text classification is an agile research area in text mining. The larger digital textual database is available in text classification for managing such data increasing. Textual data is inevitable to manage efficiently to make the access and search any precise document faster. They have certain rules are used by text classification for assign class to the unlabelled document from predefined classes. The number of tasks in text classification was solved manually

Not only in English language and also in other regional languages the use of internet exponential increase the amount of text documents is developed. For Indian languages, they have been little work done in text classification. For this reason, number of problems is deal by many Indian languages such as: not any capitalization, non-availability of abundant gazetteer list, deficiency of standardization and spelling, shortage of resources and tools, free word order language.Indian languages are highly inflectional and derivational language especially Dravidian languages such as Tamil, Telugu, Kannada, and Malayalam.[6]

In Punjabi text classification, they have not been much work done to classify the Punjabi documents due to shortage of resources and tools, name dictionaries, best morphological analyzers POS triggers are not useful in the needed measure. They have many Techniques are used to classify the text document such as: Nearest Neighbour (KNN), Bayesian classification , Support Vector Machine, Association based classification , Term Graph Model , Decision Tree , Neural Networks ,Centroid based technique etc.

IV. METHODOLOGY

1. PROBLEM DEFINITION

In problem definition, there is not any classier designed in past to classify the automatic multiple Punjabi documents. In this system, we can use centroid classifier and they have to classify automatic multiple Punjabi documents.

Now-a-days, text classification is exceedingly essential for an each range of vision to organize the text document.

Two new algorithms are proposed for Punjabi text classification, first is ontology based and second is hybrid

approach. At this point, we have certain examples of Punjabi news article which we have to classify by making use of algorithms. Punjabi is an Indian language and its spoke in west Punjab (Pakistan) and East Punjab (India). Similarly equitable influence has done in Punjabi Text Classification. The issues handled by number of Indo languages that are no capitalization of words where such intelligence is not effectual , deficiency of standardization, the act of naming the letters of a word and scarcity of instruments. Punjabi language had additional inflexion behavior than English language.

2. OBJECTIVES

The objectives of the proposed work are as follows:

- To collection of training data related to different type of classes such as Political news, Business news, sports news and entertainment.
- To development of system using centroid based technique.
- To trained the system by collected data.
- To evaluate the system.

Till now there is little work has been done in Punjabi documents for text classification. Centroid based technique is intended for Punjabi text classification. At this point, we have certain examples of Punjabi news article to classify with the help of algorithms. [7]

3. PUNJABI TEXT CLASSIFICATION ALGORITHMS

- **Centroid Based Classification:-**

The distance between each Centroid vector (c) and document vector (d) are calculated; ascribe that class to the document that is having least Euclidean distance from the Centroid vector [8]

In this technique, every document d is denoted as vector is known as Document Vector V (d) in the feature space and every component of the vector is denoted as TF*IDF value i.e. V(d)=(tf*idf1, tf*idf2,.....tf*idfn) where n is the nth word in the text document. The term "TF" stands for term frequency where a number of times the term appear in the document is known as term frequency and the term "IDF" stands for inverse document frequency .Centroid vector of every class is also computed and then Euclidean distance along document vector and centroid vector of every class are calculated. And ascribe class to the document that is having least distance from Centroid vector of precise class. [9]

Training Set	For each class document Step1: Calculate Tf. Step2: Calculate Df. Step3: Calculate Tf-IDf Step4: Store calculated values in List [ ] [ ]. Calculating centroid per class Step1: Get all words from all classes. Step2: For each word (w). Step3: For each class (c). Step4: Get Tf-IDf value for word (w) in class(c). Step5: Calculate for that word. Step6: Save value of word (w) in Array of words.
Test Set	Step1: Get all words from all classes. Step2: Calculating Tf for testing document. Step3: Create Bag of Words per class. Step4: Calculate total document. Step5: Calculate Tf-IDf. Step6: Get Euclidean per class. Step7: Calculate minimum distance from each class. Step8: Return minimum distance class name.

V. RESULTS

To evaluation of the system, we have been used testing data. Testing data has been collected from online news websites etc. Detail of testing data shown below:-

They have found out the accuracy of this system using these formulas following as:

1. Recall: $\frac{No.\ of\ correct\ output\ retuned\ by\ system}{No.\ of\ total\ files\ tested}$

2. Precision: $\frac{No.\ of\ correct\ output\ reurned\ by\ system}{No.\ of\ true\ predictions}$

3. F1-measure: $2 * \frac{R*P}{P+P}$

4. per class accuracy:

S. No.	Domain	Results
1	Business News	87%
2	Entertainment News	86%
3	Politics News	89%
4	Sports News	89%

Table 1 per class accuracy

5. Overall accuracy:

S. No.	Domain	Results
1	Recall	88%
2	Precision	89%
3	Tf-measure	89%

Table 2 Overall accuracy of the system

VI. CONCLUSION

Text classification is used to contrive the data from the set of predefined data. In Punjabi text classification, they have been little work is done. The Punjabi text is to be classifying using the statistical approach. For getting root words stemmer rules are developed. Text classification system is the vital system in the area of Natural Language Processing. Text classification is used by various online and offline system to categorize text into the set of predefined classes. The problem of classification has been widely studied in the database, data mining, and information retrieval communities. We have successfully implemented and tested Centroid-based classifier. Centroid-based classifier produces better the result than Naive Bayes classifier. The system has the capabilities to classify given text news into four different categories such as sports, business, entertainment, and politics. We are able to achieve satisfactory results based on our training data, at that moment was not available. They have various online resources helps to collect training data, which was a very challenging and time-consuming task for this system. European language like English is a resource-rich language as compared to the Punjabi language where one can find sufficient resources for training and testing the system. Based on our collected training data, which was not in a large amount , we are capable to attain satisfactory results from this classifier system. [10]

VII. REFERENCES

1) Nidhi, Vishal Gupta University Institute of Engineering and Technology, Panjab University.Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach(2012).

2) Vishal Gupta and Gurpreet Singh Lehal Department of Computer Science, Punjabi University Patiala, India. International Journal of Computer Applications (2011) .Named Entity Recognition for Punjabi Language Text Summarization.

3)Nidhi, Vishal Gupta University Institute of Engineering and Technology, Panjab University.Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach(2012).

4) Bhumika, Prof Sukhjit Singh Sehra, Prof AnandNayyar. International Journal of Application or Innovation in Engineering & Management (IJAIEM)(2013).

5) Shruti Bajaj Mangal, Dr. Vishal Goyal Research Cell : An International Journal of Engineering Sciences, Issue December (2014),Vidya Publications. Authors are responsible for any plagiarism issues. Text News Classification System using Naïve Bayes Classifier.

6) Vishal Gupta and Gurpreet Singh Lehal Department of Computer Science, Punjabi University Patiala, India. International Journal of Computer Applications(2011).Named Entity Recognition for Punjabi Language Text Summarization.

7) International Journal of Artificial Intelligence & Applications (IJAIA), March (2012) Text Classification and Classifiers:A surveyVandanaKordeSardarVallabhbhai National

Institute of Technology, SuratC NamrataMahender Department of Computer Science&IT, Aurangabad.

8) Nidhi, Vishal Gupta University Institute of Engineering and Technology, Panjab University. Domain Based Classification of Punjabi Text Documents using Ontology and Hybrid Based Approach(2012).

9) International Journal of Data Mining Techniques and Applications Indian Language Text Representation and Categorization Using Supervised Learning Algorithm. M Narayana Swamy ,M. Hanumanthappa Department of Computer Applications, Presidency College ,Bangalore ,India, Department of Computer Science & Applications, Bangalore University, Bangalore, India.

10) Bhumika, Prof Sukhjit Singh Sehra, Prof AnandNayyar. International Journal of Application or Innovation in Engineering & Management (IJAIEM)(2013).