Abstract
Text Mining is a field that extracts useful information from the text document according to users need which is not yet discovered. Text Classification is one of the text mining tasks to manage the information efficiently, by classifying the documents into classes using classification and clustering algorithms. Each text document is characterized by a set of features used in text classification method, where these features should be relevant to the task. This paper introduces preprocessing techniques, feature selection methods for classify Punjabi Text documents by clustering and classification algorithm.

Keywords
Text mining, Features Extraction, TF-IDF, Ontology Based Classification, K-Mean Clustering, Fuzzy C-Mean

I. Introduction
Text Mining is the process of extracting high-quality information from text documents ‘High quality’ in text mining means that information deriving should be relevant to the user need, and according to the user interest. The text document can be of any type plain text document or a tagged text document. And text mining tasks include clustering, classification, concept/entity extraction, topic tracking, information visualization, question answering etc. Most of the available contents are in the form of digital, and managing digital data become very difficult task [6].

Therefore, Text Classification technique is used to classify the large documents into predefined classes, this result in increased in the efficiency of the search engine.

A lot of attention has received to the field of text mining due to the increasing need for managing the information that is available in the vast amount of documents. In text classification documents are classified with supervised knowledge. Text Classification task is two-step processes [7]

A. Training Phase
The set of documents in training phase is known as training set. The training sets contain documents that belong to particular class based on their contents, called Labeled Documents. With the help of Training Phase helps we can classify the unlabeled documents.

B. Testing Phase (Classification Phase)
This phase classify, unlabeled documents using labeled documents. Features contain the important information about the contents of the document. Feature selection is the task to identify a subset of relevant features from a document, so as to achieve classification accuracy. The clustering process is a general approach on the properties that data share. Three major types of clustering processes according to the way they organize data are hierarchical, partitioning and mixture model methods [9]. If the data is clustered according to the property of data, its character and the behavior, then cluster impact is consider to be valuable. Two different types of clustering algorithms are [9].

C. K-Means Algorithm
K-Means is unsupervised learning algorithms that solve the clustering problem. The procedure follows classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main aim is to define k centroids, one for each cluster. Because of different location causes different result these centroids should be placed in a cunning way. So, to place them as much as possible far away from each other is the better choice. Next step is to take each point belonging to a given data set and associate it to the nearest centroid.

D. Fuzzy C-Means Algorithm
Fuzzy C-Means (FCM) is an unsupervised clustering algorithm. This is applied to wide range of problems involving feature analysis, clustering and classifier design. FCM works over wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target recognition. The clusters are formed according to the distance between data points and cluster centers are formed for each cluster. The Algorithm Fuzzy C Means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters.

Punjabi text classification process divides into three phases [1].

E. Preprocessing Phase
This phase include process such as, removal of stop words, stemming, punctuation mark and special symbols removal.

F. Feature Extraction
This phase includes statistical approach and linguistic approach for the extraction of relevant features from the documents to perform classification.

G. Processing Phase
The last phase of the Punjabi text classification, apply text classification algorithms to the extracted features to classify the documents into classes.

II. Related Work
Text Classification is an active research area of text mining, as there is a large digital textual database available, therefore demand of managing such data increasing. It becomes necessary to manage this textual data efficiently to make the access, search of any particular document faster [5].

Text Classification uses certain rules to assigns class to the unlabelled text documents from a set of predefined classes. Many text classification tasks was solved manually, Traditionally. But such way of classification is expensive and also labor intensive since rules are created manually for the efficient classification. Therefore, for the betterment another approach of classification is machine learning-based text classification that uses automatic generation of rules to classify the documents [7].

Techniques for the classification are: Nearest Neighbour (KNN) [7], Bayesian classification [7], Support Vector Machine [7], Association based classification [7], Term Graph Model [7],
Decision Tree [7], Neural Networks [7] etc. The development and use of internet exponential increase the amount of electronic documents not only in English language but also in other regional languages. Very little work has been done for text classification with respect to Indian languages. So far this is because of problems faced by many Indian Languages such as: No capitalization, non-availability of large gazetteer lists, lack of standardization and spelling, scarcity of resources and tools, free word order language. Indian Languages, especially Dravidian languages (Tamil, Telugu, Kannada, and Malayalam) are highly inflectional and derivational language, leading to a very large number of word forms for each root word. This thing makes the classification task more challenging.

But for Punjabi Text Document, not much work has been done to classify the documents due to lack of resources, annotated corpora, name dictionaries, good morphological analyzers, POS taggers are not yet available in the required measure. Therefore, in section 3, we present a classification algorithm for Punjabi Text Classification.

III. Preprocessing Phase

We represent each original text document as “Bag of words” [1]. Then following operations are done on each document:

A. Remove Punctuation Marks

Remove punctuation marks, special symbols (., .., ::, ., ^, &, #, etc.) from the Punjabi text documents. And also, if a Punjabi text document contains excess use of spaces, tabs, shift, then remove them.

B. Stop Words Removal

Stop words Removal, Stop words are the set of words that occur very frequently. Which do not contain any useful or relevant information to the text classification task. Therefore, for the efficient working of the text classification we have to remove these words from the text documents. We have made a list of Punjabi language stop words from the dataset. This corpus contains around 47,029 words. Some commonly occurring stop words are: ਹੈ (dē), ਹੋ (dā), ਸ੊ੱਚ (vīc), ਹੋ (dī), ਹਦ (hai), ਹਕਰੀ (ih), ਅਖ਼ (atē), ਬਚਾ (valōp), ਹਤਾ (han), ਹੁਣ (nūp) etc.

C. Stemming

Stemming In this process we reduce the inflected (or sometimes derived) words to their root form. An analysis of corpus was made and various possible noun suffixes were identified like ਨਾਂ (tān), ਹੋ (tān), ਸੋ (tē), ਧਿਆਂ (tēn)) etc. its programming language dependent. Therefore, for large dataset, like we have 150 Punjabi text documents, we use Punjabi Language Stemmer given in. Input: ਕਲੀ (kēlē); Output: ਕਲੀ (kēlē).

After preprocessing phase, words than at contain relevant information to the classification tasks are left. Now, we have the documents which contain less numbers of words than the original documents, and extracting features from these words now become easier.

IV. Feature Extraction

Features extracted is consider to be a important part of the text classification if we carefully chose the features it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

Due to the high dimensionality of feature sets, feature extraction can be performed to improve the efficiency and accuracy of the classifiers. Better understanding about the contents of the documents is done by Feature extraction and also provide information about important features and how they are related with each other. A good feature set contains information how one object is distinguish from other objects. The selected set of features should be a small set [8].

There are two approaches to extract features from the text document that are following:

A. Statistical Approach

These methods are simple, language independent and do not require training data. Following are the different methods [5] [1].

1. Term Frequency Weighting (TF)

In this method, to the number of times the term appear in the document is known as the weight of a term in a document, i.e. the frequency of the term in the document.

\[ W_i = tf_i = tf_i \cdot \log \left( \frac{N}{n_i} \right) \]  

where \( i \) is the index of the term, \( N \) is the total number of words in the document, \( n_i \) is the number of times the term \( i \) appears in the document.

2. Term Frequency-Inverse Document Frequency Weighting (TF-IDF)

TF weighting do not consider the frequency of the term throughout all the documents in the document corpus. Term Frequency * Inverse Document Frequency (TF *IDF) weighting is the most common method used for term weighting that takes into account this property. In this approach, the weight of term \( i \) in document \( d \) assigned proportionally to the number of times the term appears in the document, and in inverse proportion to the number of documents in the corpus in which the term appears.

\[ W_i = tf_i \cdot \log \left( \frac{N}{n_i} \right) \]  

TF *IDF Weighting approach weights the frequency of a term in a document with a factor that discounts its importance if it appears in most of the documents, as in this case the term is assumed to have little discriminating power.

3. TF *IDF Weighting with Length Normalization

In this approach, to account for documents of different lengths each document vector is normalized so that it is of unit length.

\[ W_i = \frac{tf_i \cdot \log \left( \frac{N}{n_i} \right)}{\sum_{j=1}^{N} \left[ tf_j \cdot \log \left( \frac{N}{n_j} \right) \right]^2} \]  

This method performs better than other methods.

4. Mutual Information (MI)

This statistical method is used to determine whether a genuine association exists between two text features or not. In text classification, MI has been broadly employed in a variety of approaches to select the most significant text-features that serve to classify documents. MI between term \( t \) and class \( c \) is calculated by:
\[ M(t,c) = \log \frac{A_{ct} \times N}{(A_{ct} + C_c) \times (A_{ct} + B_c)} \]  

Here, \( A_{ct} \) is the number of times \( t \) and \( c \) co-occur, \( C_c \) is the number of times \( t \) occurs without \( c \), \( B_c \) is the number of times \( c \) occurs without \( t \), and \( N \) is the total number of documents. When \( t \) and \( c \) are independent \( M(t,c) \) is equal to zero.

### B. Linguistic Approach

A number of rules specific for Punjabi language are formed to extract the language dependent features [1].

#### 1. Name Rules

(i). **Person-Prefix**

If current word is prefix, its next word is taken as First Name.

(ii). **Middle Name and Last Name**

- Word next to first name is checked for middle name or last name.
- If it is middle name, its next word is checked whether it is last name or not.
- If not, middle name is considered as last name.

#### 2. Location Rules

- If Punjabi word विखेल (vikhē) is found, its previous word is extracted as location name.
- If Punjabi word पिंड (piṇḍ) is found, its next word is extracted as location name.
- If Punjabi word जिलहेल (zilhē) is found, its previous word is extracted as location name.

#### 3. Date/Time Rule

- If month is found, it is extracted.
- If week day is found, it is extracted.
- If Punjabi words दिन (ik), दिन (dūjā), दिन (dō), दिन (pahilā), चेविं (chēvī) etc. are found, they are extracted.

#### 4. Numbers/Counting

If any numeric character is found, it is extracted. If Punjabi words इक (ikk), दो (dūjā), तीन (tīn), पाँच (pāch), चार (cār) etc. are found, they are extracted.

#### 5. Designation Rule

If designation name found e.g. कप्तान (kaptān), कॉक (kōc), कैप्टन (kaiptān), it is extracted.

#### 6. Abbreviation

If words like आई (āī), उ (ū), आ (āil), ओ (pī), बी (bī) etc. are found, they are extracted.

### V. Proposed Classification

**Algorithm**

Step1: Collect the text from the Punjabi documents (like sports, matrimonial and politics etc.)

Step2: Remove all additional symbols, extra tabs, spaces, stopwords, shifts from the text documents.

Step3: Extract names, places, dates, months name etc the text document using Feature Extraction.

Step4: Calculate Term Frequency (TF) for each remaining word and eliminate terms whose term frequency is below the threshold value.

Step5: Calculate Inverse Document Frequency each word from the document after pre step.

Step6: Calculate TF * IDF of each word those words that are having TF less than threshold value. This step will further help in reducing dimensionality.

Step7: Use the Data mining technique to create clusters of different ontologies.

Step8: Create ontology for each class that consists of terms related to its classes e.g. for Cricket Class and SGPC Class Ontology for sports and Politics, we have terms such as गेंदबाजी (gēndbāzi), विकाट (vıkāṭ), सापिन (sapīn) (vıkṭakīpar), (gurudwara parbandhak) etc. This results in Class wise list.

Step9: Remaining terms from step 8 is matched with each Class-wise list, and if maximum terms are matched with one class, assign that class to the unlabelled document.

### VI. Dataset

Documents are taken from the Punjabi news web sources such as likhari.org, jagbani.com, ajitweekly.com, punjabispectrum.com, europevichpunjabi.com, quamiekta.com, sahitkar.com, onlineindian.com, europesamachar.com, parvasi.com etc.

**Fig. 1: Algorithm for the Punjabi Text Document**

### VII. Conclusion

Extraction of text from Punjabi news paper literature is an essential operation. Given that there have been many text extraction methods developed this paper presents a novel technique that employs classification based article clustering to further get the text extraction process. The development of the proposed work is of practical significance; however it is challenging to design a unified approach of text extraction that retrieves the relevant text articles of Punjabi newspapers more efficiently. Collect the Documents from the Punjabi news web sources such as likhari.org, jagbani.com, ajitweekly.com, punjabispectrum.com etc. The proposed algorithm, using data mining technique called text-mining, seems to classify the text of seven keywords i.e. sports...
section: lawn tennis, cricket, hockey. Politics section: National and centre level, Matrimonial section etc. Create the different ontologies for the Punjabi text. In this way the classification of Punjabi documents can done.

References


Shruti Aggarwal did her Master’s in Engineering in Computer Science from U.I.E.T., Panjab University, Chandigarh in 2011. She did her B.Tech (Computer Science & Engineering) from Ambala College of Engineering and Applied Research, Kurukshetra University in 2008. She has teaching experience of one year as lecturer at Emax Group of Institutes, Ambala. She has also taught at Dr. B.R.Ambedkar National Institute of Technology, Jalandhar for one semester. From last one year, she is working as Assistant Professor at Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India. Her main research area is Data Mining and she has published many research papers in International Journals and Conferences on Audio Mining and other Data Mining Techniques.