Review on Knowledge Extraction Using Classifier
Nirmal Singh¹ Shruti Aggarwal²
¹M.Tech, Research Scholar ²Assistant Professor
¹,²Department of Computer science Engineering
¹,²Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

Abstract— Text mining is a process of extracting knowledge from large text documents. There are various methods of text mining that are classification, clustering, association and regression. Classification is finding models that analyze and classify a data item into several predefined classes. Models with the help of which data items are classified called classifiers. There are various types of classifiers. Documents are grouped in such a way that there may be the fast retrieval. The main task of classification is mapping a data item into one of several predefined classes.

Key words: supervised, unsupervised semi supervised, NLP

I. INTRODUCTION
Provident Fund is the fund which is composed of the with increasing number of the electronic documents from a variety of sources text mining gets more importance. It includes unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, Information retrieval, classification (supervised, unsupervised and semi supervised) and summarization. Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents [1].

Text classification is an important part of text mining. The main task of this is to determine classical model which is used to assign class labels to the documents. On the basis of documents, classification can be categorized as single label and multi label. Single label document is belongs to only one class and multi label document may be belong to more than one classes.

II. THE CLASSIFICATION PROCESS
There are various steps in classification:
A. Document Collection
In this document of different formats are collected. It may be pdf file or html file etc.
B. Pre-Processing
The first step of pre-processing which is used to presents the text documents into clear word format. It has various steps:
1) Tokenization
A document is treated as a string, and then partitioned into a list of tokens.
2) Removing stop words
Stop words such as “the”, “a”, “and”, etc are frequently occurring, so the insignificant words need to be removed.
3) Stemming word
Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of reducing tokens to their root form, e.g. connection to connect, computing to compute.

C. Indexing
The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector. The most commonly used document representation is called vector space model. The vector space model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document Matrix.

Vector Space Model have some limitations that are loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document. To overcome these problems, term weighting methods are used to assign appropriate weights to the term. But the main drawback is high dimensionality.

D. Feature Selection
After pre-processing and indexing the important step of text classification, is feature selection [2] to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space. Many feature evaluation metrics have been notable among which are Information gain (IG), term frequency, Chi-square, expected cross entropy, Gini index etc.

E. Classification
The automatic classification of documents into predefined categories is an active attention. The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. There are different classifiers for classification that are Bayesian classifier, Decision Tree, K-nearest neighbor (KNN), Support Vector Machines (SVMs), Neural Networks, used to classify data.

III. CLASSIFIERS
A. Rocchio’s Algorithm
Rocchio’s learning algorithm [4] is in the classical IR tradition. It was designed to use relevance feedback in querying databases. It is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

B. K-Nearest Neighbors
K-NN classifier is a case-based learning [5] algorithm that is based on a distance or similarity function for pairs of
observations, such as the Euclidean distance or Cosine similarity measure’s.

This method is used in many applications [6] because it is effective, non-parametric and easy to implementation properties, however the classification time is long and difficult to find optimal value of k. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make less distinct boundaries. To overcome these limitations replace the traditional KNN with different K-values for different classes rather than fixed value for all classes.

C. Naïve Bayes

Naïve bias method is kind of module classifier [7] under known priori probability and class conditional probability. It is basic idea is to calculate the probability that document D is belongs to class C. There are two event model are present for naive Bias [8] [9] [10] which are multivariate Bernoulli and multinomial model. Out of these, model multinomial model is more suitable when database is large but multinomial model have some problems. As it have rough parameters and can handle only few training documents. Naïve Bayes is easy for implementation and computation.

D. Decision tree

When decision tree is used for text classification it consist of internal node are label by term, branches produced from them are labeled by test on the weight, and leaf node are represent corresponding class labels.

Tree can classify the document by passing from root to leaf nodes which represents the goal of classification of the document. Due to swapping of training tuples, most of the data will not fit in memory. To handle this issue [11] presents method which can handle numeric and categorical data.

E. Decision Rule

This method uses the rule based inference to classify documents it uses disjointive normal form model[12]. When data set is large no of features are implemented to reduce the size of rules set without affecting the performance of classification.

F. SVM

The application of Support vector machine (SVM) method to Text Classification has been proposed by [13]. The SVM need both positive and negative training set which are not used for other classification methods. These positive and negative training set are needed for the SVM that are used at decision surface that best separates the positive from the negative data in the m dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. SVM classifier is better because of its effectiveness to improve text classification. It is capable to solve the multi label class classification.

G. Neural Network

Neural network classifier is a network of units where input unit represent terms and output unit represents category. To classify a document weights are assigned to input unit this unit is activated through network and the value of output unit determines the categorization decision. An efficient feature selection method [14] is used to reduce the dimensionality as well as improve the performance. New Neural network based document classification method. [15] Was presented, which is helpful for companies to manage patent documents more effectively.

H. LLSF

LLSF stands for Linear Least Squares Fit in which training data is represented in the form of input and output vector pairs. Input vector is a document in vector space model and output vector consists of categories that are binary weights. There is a limitation of it that the computational cost of computing the matrix is much higher.

I. Voting

This algorithm is based on method of classifier committees and is based on idea that given task that requires expert opinion knowledge to be performed. Different combination rules are present as the simplest possible rule is majority voting (MV) If two or three classifiers are agree on a class for a test document, the result of voting classifier is that class. Second weighted majority voting, in this method, the weights are specific for each class in this weighting method, error of each classifier is calculated.

J. Associative classifier

Associative classifiers is powerful and achieve high accuracy. The main idea behind this algorithm is to scan the transactional database searching for k-item sets relationships among items in a transactional database.

K. Centroid based classifier

Centroid based classifier is simple for each set of documents belonging to the same class, we calculate their centroid vectors. If there are k classes in the training set, this leads to k centroid vectors (C1, C2, C3...) where each Cn is the centroid for the jet class. The class of a new document x is determined as, First the document-frequencies of the various terms computed from the training set Then, compute the similarity between x to all k centroid using the cosine measure. Finally, based on these similarities, and assign x to the class corresponding to the most similar centroid.

IV. KNOWLEDGE EXTRACTION USING PROBABILISTIC CLASSIFIER

Recently a new classifier is proposed called probabilistic framework for information extraction. In this graphical model using conditional probability is used for relational data. It has three steps:

A. Preprocessing

In this the input document is split into paragraphs and statements. Stop words and pure words are identified and removed.

B. Rule Generation

In this document from training set is read, then positive and negative rules are generated.

C. Probability Calculation

We compute weight factor for each category. Positive weight factor pw is calculated between the number of positive rules (Npr) for the terms in the document toward a category the total number of terms in the document. Negative weight factor nw is calculated using the Number of negative rules (Nnr) for the terms in the document toward a category and total number of terms in the document.

After calculating weight factors probability is calculated then positive probability is determined.
V. ENHANCEMENT

For text mining there are various algorithms but, for enhancement for the purpose of classification the probabilistic classifier can be used. This classifier may produce better results when positive and negative probability of a particular category is closely determined means strong rules are required to implement in order to modify and refine the results. There are various classifier discussed above out of all probabilistic classifier produce more general results, so needs to be enhanced. When classification is performed it can be analyzed by various algorithms.

VI. CONCLUSION

The growing use of the textual data that needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. The discussion about existing literature and explanation of documents representation and an analysis of feature selection methods and classification algorithms were presented but, it is understood that no single representation scheme and classifier can be mentioned as a general model for any application. Different algorithms perform differently depending on data collection. However, to the certain extent SVM with term weighted VSM representation scheme performs well in many text classification tasks. Out of these it can be examined that probabilistic classifier when modified can produce precise results because it depends on probability factor of any category.

REFERENCES

[14] Hwee TOU Ng Wei Boon Goh Kok Leong Low, “Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization”, SIGIR 97 Philadelphia PA.