JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text Classification using Ontology Graph Representation through Bag of Words

[1] Kavya S N , [2]Dr Sudhamani M ,[3]Jagadeesh M and [4]Abhishek M Anegundi
[1] Department of Computer Science, MMK & SDM MMV, Mysuru.
[2] Department of Computer Science, MMK & SDM MMV Mysuru,
[3]Department of Computer Science, University of Mysore.
[4]Department of Computer Science, University of Mysore.

Abstract
Text classification is the process of automatically categorizing text documents into a set of predefined classes. Ontology deals with similar terms and relationship that can be used to describe and represent area of knowledge. In this work, we categorize text to a given ontology by using the Knowledge of the document, knowledge represented in the form of ontology for categorizing documents. We experimented the on 20 newsgroup mini dataset, 20 newsgroup large dataset and Reuters-21578 datasets. Support Vector Machine (SVM) classifier is used to classify the text documents and performance of the classifier is measured in terms of accuracy and f-measures.

Keywords: Text categorization, ontology, Support Vector Machine (SVM), knowledge representation, feature extraction.

1. Introduction
Text classification methods are mainly used to determine the belonging of text document. The purpose of classification is to identifying the belongingness of text. Rapid development of information technology directly affects an increase in the amount of data. Automatic text classification methods enable the organization or categorization of a set of documents into different categories or classes. Many classification problems have been solved manually by the use of some rules commonly written by hand. Instead of using hand-written rules, the text categorization approaches use machine learning methods to learn automatic classification rules based on human labelled documents. It is obvious labelling is easier than defining rules. The rapid growth in data volume increases the complexity of classification; it also requires a lot of time and human effort for manual classification. Therefore, automated classification of the electronic documents needed. Hence, text categorization can be considered as an effective method for automatic assignment of documents to the predefined categories according to their context. In this paper, we consider the tasks of the automated classification of 20NewsgroupLarge dataset, 20 Newsgroup Mini dataset,REUTERS-21578(Transcriptions).

Text classification is the process of assigning tags or categories to text according to its content. It's one of the fundamental tasks in **Natural Language Processing** (NLP) with broad applications such as sentiment analysis, topic labelling, spam detection, and intent detection. Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes.
It is observed that information extracted is enormous, adequate and are unstructured in nature. Automatically retrieving such huge information from a database could be one of the most difficult tasks in the field of information retrieval. In machine learning and text processing, Text classification plays an important role in deciding the label for the document. Much research work has been done in single label text classification problems, where single-labelled document means a document belongs to single class. Few works have done in

multi-label classification problem, if a document belongs to more than one class in the corpus, then the document is called as a multi-labelled document.
[1] Frequently text document structuring can be either done by either by ontologies and metadata or by automatic unsupervised text categorization.
Further the paper will discuss the existing work in related work section, proposed model is discussed in proposed methodology, experimental results will be considered in experimental analysis and we conclude the work and convey the future work in conclusion and future work.

2. Related Work

[1] The integrated framework OTTO (Ontology-based Text mining framework). OTTO uses text mining to learn the target ontology from text documents and uses the same target ontology in order to improve the effectiveness of both supervised and unsupervised text categorization approach.

[2] A text mining method using ontology based hybrid approach for classification and clustering of research project proposals based on the keyword of different disciplines. It provides mining of text and optimization to improve the proposal grouping process based on its similarities.

[3] A framework on ontology based text mining for grouping research papers and assigning the grouped paper to reviewers systematically. Research ontology is constructed to categorize the concept terms in different areas and to form relationships among them. It facilitates text-mining and optimization techniques to cluster research papers based on their similarities and then to assign them to reviewer according to their concerned research area.

[4] Huang Y et al proposed an innovative text categorization model, VSM_WN_TM, based on Vector Space model (VSM), Support Vector Machine is used as document classifier and the proposed system is evaluated on publicly available datasets and domain-specific dataset. Experiment result shows that incorporating semantic and syntactic relationship among words such as synonymy, co-occurrence and context could greatly improve text representation and the method significantly outperforms conventional approaches such as using only BOW features or latent topic features.

[5] Sahri et al, team aim to represent and understand the financial criminology domain in the form of Ontology Knowledge Representation. The domain representation helps researcher and investigator to understand the area of study and key issues in financial criminology. Industry people are able to use the ontology to develop a knowledge base system of financial crime. In this research data extracted from 25 research papers on Financial Criminology gained from main online journal databases such as Scopus and Web of Knowledge (WoS). Then the research uses the terms and words to design the ontology based on the methods explained in the methodology section. The result found that there are nine main classes of common research on financial criminology are namely; People, Location, Time, Property, Offenses, Risk Sector, Enforcement, Technology and Resources.

[1] BloehdornS. Et al combined natural language processing techniques and machine learning algorithms to construct ontology's in semi-automatic manner and the method is known as ontology learning. Ontology learning is critical for building domain specific ontology's with fewer manual efforts. Background knowledge in form of ontology's enhances the performance of classical text mining tasks such as text classification and text mining tasks such as text classification and text clustering. Semantic features extracted from ontology's with help of the OTTO text mining components leverage the classical bag-of-words representation to a higher semantic level and thereby improve classification accuracy and cluster purity.

[8] The steady accumulation of unstructured data, the domain of natural language processing (NLP) is gaining widespread attraction amongst researchers and practitioners in order to quickly and easily extracts prediction-like insights in a simplified and streamlined fashion. There are a number of articulations on the subjects of NLP and machine learning (ML). Very recently, the model of the bag of words has become so popular in order to produce accurate predictions out of unstructured text data. In this paper, they explained an easy-to-use framework for accelerated usage of the BoW model towards pioneering text mining and processing and demonstrated a simple example by leveraging the framework in order to showcase the utility of this generic framework that can be easily replicated across in many other associated scenarios.

[14]In this paper, all the applied methods on feature extraction on text categorization from the traditional bag-of-words model approach to the unconventional neural networks are discussed.

[15] Hybrid methods are very important for feature selection in case of the classification of high-dimensional datasets. In this paper, we proposed two hybrid methods which are the combination of filter-based feature selection, genetic algorithm, and sequential random search methods. The first proposed method is hybridisation of information gain and genetic algorithm. In this, first, the features are ranked based on the information gain and then a user defined features are selected from the ranked features. Genetic algorithm with these selected features is applied for the selection of optimal feature subset. It is applied for feature selection with two types of fitness functions which are single objective and multi-objective in nature. The second feature selection model is the hybridisation of information gain and sequential random K-nearest neighbour (SRKNN). In this method, again information gain is used to rank the features and a user defined top ranked number of features are selected. A set of binary population (having all feature selected by users) are generated and on each population sequential search method is applied for maximising the classification accuracy. These methods are applied to 21 high-dimensional multi-class datasets. Obtained results show that on some datasets first method's performance is good and on some datasets second method's performance is good. The results obtained by proposed methods are compared with results registered for other methods.

[16] Adequate selection of features may improve accuracy and efficiency of classifier methods. There are two main approaches for feature selection: wrapper methods, in which the features are selected using the classifier, and filter methods, in which the selection of features is independent of the classifier used. Although the wrapper approach may obtain better performances, it requires greater computational resources. For this reason, lately a new paradigm, hybrid approach, that combines both filter and wrapper methods has emerged. One of its problems is to select the filter method that gives the best relevance index for each case, and this is not an easy to solve question. Different approaches to relevance evaluation lead to a large number of indices for ranking and selection. In this paper, several filter methods are applied over artificial data sets with different number of relevant features, level of noise in the output, interaction between features and increasing number of samples. The results obtained for the four filters studied (Relief, Correlation based Feature Selection, Fast Correlated Based Filter and INTERACT) are compared and discussed. The final aim of this study is to select a filter to construct a hybrid method for feature selection.

[17]Text classification is the process of assignment of unclassified text to appropriate classes based on their content. The most prevalent representation for text classification is the bag of words vector. In this representation, the words that appear in documents often have multiple morphological structures, grammatical forms. In most cases, this morphological variant of words belongs to the same category. In the first part of this paper, anew stemming algorithm was developed in which each term of a given document is represented by its root. In the second part, a comparative study is conducted of the impact of two stemming algorithms namely Khoja's stemmer and our new stemmer (referred to hereafter by origin-stemmer) on Arabic text classification. This investigation was carried out using chi-square as a feature of selection to reduce the dimensionality of the feature space and decision tree classifier. In order to evaluate the performance of the classifier, this study used a corpus that consists of 5070 documents independently classified into six categories: sport, entertainment, business, Middle East, switch and world on WEKA toolkit. The recall, f-measure and precision measures are used to compare the performance of the obtained models. The experimental results show that text classification using rout stemmer outperforms classification using Khoja's stemmer. The f-measure was 92.9% in sport category and 89.1% in business category.

[18]Text mining, also known as Intelligent Text Analysis is an important research area. It is very difficult to focus on the most appropriate information due to the high dimensionality of data. Feature Extraction is one of the important techniques in data reduction to discover the most important features. Processing massive amount of data stored in a unstructured form is a challenging task. Several pre-processing methods and algorithms are needed to extract useful features from huge amount of data. The survey covers different text summarization, classification, clustering methods to discover useful features and also discovering query facets which are multiple groups of words or phrases that explain and summarize the content covered by a query thereby reducing time taken by the user.

[19]As the world is moving towards globalization, digitization of text has been escalating a lot and the need to organize, categorize and classify text has become obligatory. Disorganization or little categorization and sorting of text may result in dawdling response time of information retrieval. There has been the 'curse of dimensionality' problem, namely the inherent sparsity of high dimensional spaces. Thus, the search for a

possible presence of some unspecified structure in such a high dimensional space can be difficult. This is the task of feature reduction methods. They obtain the most relevant information from the original data and represent the information in a lower dimensionality space. In this paper, all the applied methods on feature extraction on text categorization from the traditional bag-of-words model approach to the unconventional neural networks are discussed.

General Terms Text mining, feature extraction, neural networks, deep learning

[20]During the last few years there is a remarkable increase in development of Data mining techniques. Nowadays, various organizations store their information as a kind of databases. So huge amount of data and their information are available in repositories. To interpret these data, we need effective mining techniques for better performance of classifications. This helps us to take best decision and prediction. This paper is prepared to analyze the performance of classification algorithms with help of data mining tool TANAGRA and Rough set theory. At the same time, we can reveal the best tool among them based on their performance level. To experiment this, huge size data has taken which tells that performance of classification tools are affected by the kind of dataset and significant results are discussed over comparative analysis

3. Proposed Methodology

Text classification is the process of assigning tags or categories to text according to its content. Document tagging is the basic task Natural Language Processing(NLP) with broad applications such as sentiment analysis, topic labelling, spam detection, and intent detection. Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes. A typical multi-label text classification system involves the following steps:

(1) Pool of documents
(2) Pre-processing
(3) Text representation
(4) Classification/clustering
(5) Evaluation /Validation

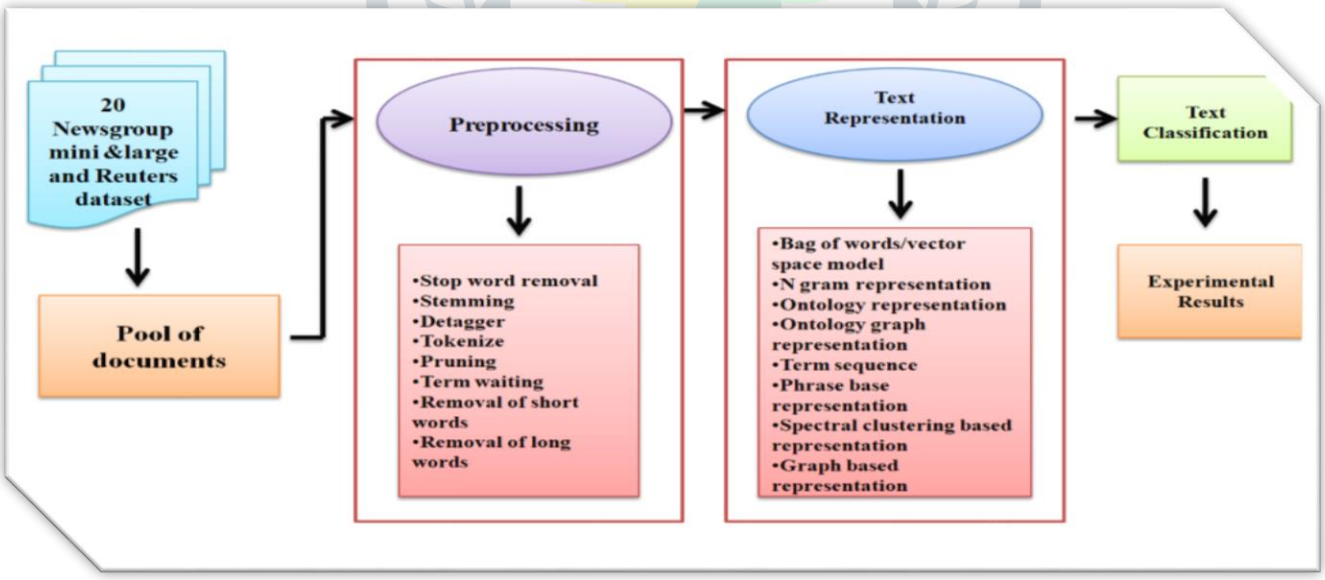


Fig.3. Block diagram of Text Classification system

3.1 Pool of Documents

It is a first stage of a text classification model concerned collecting documents of various formats belonging to the different domains. Since the collected documents are of various formats and unstructured in nature, we need to transform them into a structured presentation model. Hence, we need to employ pre-processing techniques on the unstructured and semi-structured data. Some of the widely used pre-processing techniques are: Collection reader, Detagger, Tokenizer, Stop word removal, Stemming.

3.2 Pre-processing

Pre-processing techniques are used to convert the unstructured and semi-structured data into a structured format. Some of the widely used pre-processing techniques are Collection reader, Detagger, Tokenizer, Stop word removal and Stemming. Any of the above mentioned pre-processing technique can be employed to convert unstructured textual documents into a structured format.

- **Detagger**: Is a dual-purpose tool where it converts HTML to text, or selectively removes HTML markups and removes punctuations such as “,”,“”.
- **Tokenizer**: Is used to break a string down into a stream of terms or tokens. It identifies nonempty sequences of characters, excluding spaces and punctuations in documents.
- **Stopword removal:** Removes very commonly used in a given documents like an, a, the, how, to, are, that, at, etc.
- **Stemming:** Stemming is the process of reducing inflected words to its stem, base or root form, generally a written word form.
  Example: automatic and automation are reduced to automat.
- **Pruning:** Pruning eliminates the words appearing rarely or more frequently in a document..
- **Term waiting**: It is a process of assigning weights to each term based on its frequency or relevance.
- **Removal of short words**: It removes the short words in the entire document.
- **Removal of Long words**: It is a process of removing long words from all the document. .

3.3 Text representation:

Text documents can either be in semi-structured form or in unstructured form. Converting the text documents from an unstructured format into a structured text format requires designing an ideal text representation model.

3.3.1 Bag of Words or Vector Space Model Representation

The most basic form of text representation available in literature is binary representation. The importance of a specific word is represented by 0 and 1. A value ‘1’ represents the fact that the corresponding word is extracted from the document and a value ‘0’ represent that the corresponding feature is not contained in the document. Such a feature based document representation will lads a huge sparse matrix construction because, all the words in thae documents are not equally important. Hence, replacing the binary term feature with its weight be a one probable solution. Such a weight μbased document representation is known as bag of words representation. In this model, the vocabulary defines a higher dimensional Vector space and the documents are represented as vector, where the value of the j^{th} component of the i^{th} vector is the weight of the j^{th} word in the i^{th} document. This representation is often called vector space model (VSM).

Vector space model is an algebraic model for representing text documents as vectors of identifiers, such as index terms. The vector space model represents documents as vectors in a t-dimensional space. Where, each documented is described by a numerical feature vector $d(w)=(w_1,w_2,\ldots\ldots w_n)$. Each element of the vector usually represents a word of the document collection. The size of the vector is defined by the number of words in the document. Dimension of the vector space is t, where t is the number of distinct terms, Term vectors are pair wise orthogonal, and w_{it} are the components of i^{th} document along the direction of i^{th} term. To improve the performance more weights are assigned to frequently used terms in the relevant documents but rarely used in the document pool.

3.3.2. N Gram Based Representation

An N-gram is a sub-sequences of N items from a given sequences. An N-gram is an N-character slice of a language string. In this method we need to append the blanks to the beginning and ending of the string in order to help with matching the beginning- of-word ending-of-word situations Method of generating N-gram frequency profiles for documents can be found in (Bekkerman and Allan, 2003).

An efficient text categorization algorithm that generates bigrams (N is 2) was proposed by (Tan et al., 2002). The algorithm uses the information gain metric combined with various frequency thresholds. The bigrams, along with unigrams (N is 1) are given as features to a classifier.

3.3.3. Ontology Representation

Ontology is an explicit specification of a conceptualization. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationship among them, are reflected in the representational vocabulary, pragmatically, a common ontology defines the vocabulary with which queries and assertion are exchanged

among agents. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. A commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertion using the vocabulary defined in the ontology.Ontology facilitates to capture and construction of domain knowledge and enables representation of skeletal knowledge to facilitate integration of knowledge bases irrespective of the heterogeneity of knowledge sources.

Design criteria for ontologies:

- **Clarity**: Ontology should effectively communicate the intended meaning of defined terms. Definition should be objective.
- **Coherence:** Ontology should be coherent. That is, it should allow inferences that are consistent with the definition
- **Extendibility**: Ontology should be designed to anticipate the uses of the shared vocabulary. It should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology monotonically.
- **Minimal encoding bias**: The conceptualization should be specified at the knowledge level without depending on a particular symbol-level encoding.
- **Minimal ontological commitment**: Ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities.

3.3.4. Ontology Graph Representation

A conceptual description of ontology including concepts, attribute, entity, association description and main purpose for knowledge sharing and reuse.

3.3.4.1. Classification of Ontology

Ontology is broadly classified into three categories namely, upper ontology, Domain ontology, Hybrid ontology.

- **Upper ontology:** An upper ontology is a model of the common relationship and objects that are generally applicable across a wide range of domain ontologies. It usually depends on glossary that contains terms and associated object description as they are used in various domain ontology.
- **Domain ontology:** It represent concept of word, such as biology or politics. Each domain ontology typically models domain specific definition of terms.
  Eg : The word **'card'** has many different meanings. Ontology about the domain of **'poker'** would model the **"playing card"** meaning of the word. While an ontology about the domain of **'computer hardware'** would model the **"punched card"** and **"video card"** meanings.
- **Hybrid ontology:** The Gelish ontology is an example of a combination of an upper and domain ontology.
-

3.3.4.2. Components

Current ontologies share many structural similarities, regardless of the languages in which they are expressed. Common components of ontologies include

- **Individual**: Instances or objects.
- **Classes:** Sets, Collections, Concepts, Classes in programming, Types of objects.
- **Attributes:** Aspects, Properties, Features, Characteristics or parameter that objects can have.
- **Relations**: The way in which classes and individuals can be related to one another.
- **Function terms:** Complex structure from certain relations that can be used in place of an individual term in a statement.
- **Restrictions:** Formally stated description of what must be true in order for some assertion to be accepted as input.
- **Rules**: Statement in the form of an if then sentence that describe the logical inferences that can be drawn from an assertion in particular form.
- **Axioms:** Assertion in a logical form that together comprise the overall theory that the ontology describes in its domain application.
- **Events:** The changing of attributes or events.

3.3.4.3. Ontology languages

Ontology languages are formal languages used to construct ontologies. They allow the encoding of knowledge about specific domains and frequently reasoning rules that support the processing of that knowledge. Ontology languages are usually declarative languages are almost always generalization of frame languages and are commonly based on either first order logic or on description logic.

E.g.: - Word Net, Web Ontology languages, Basic Formal Ontology.

3.3.4.4. TF-IDF

TF-IDF stands for term frequency-inverse document frequency. TF-IDF weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

- **Term Frequency (TF)**: TF is a scoring of the frequency of the word in the current document. Since every document is of different length, it is possible that a term would appear much more times in lengthier documents than shorter ones. The term frequency is calculated using the eq(1).

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ trems\ in\ the\ document} \quad \ldots \text{ eq (1)}$$

- **Inverse Document Frequency (IDF)**: IDF is a scoring of how rare the word is across documents. IDF is a measure of how rare a term is. Rarer the term, more is the IDF score. IDF is computed using eq (2).

$$IDF(t) = \log\left(\frac{Total\ number\ of\ Documents}{Number\ of\ documents\ with\ terms\ t\ in\ it}\right) \ldots. \text{Eq (2)}$$

Thus, $TF - IDF\ score = TF * IDF$ ..... Eq (3)

3.3.5. Ontology Graph Representation Through Bag Of Words

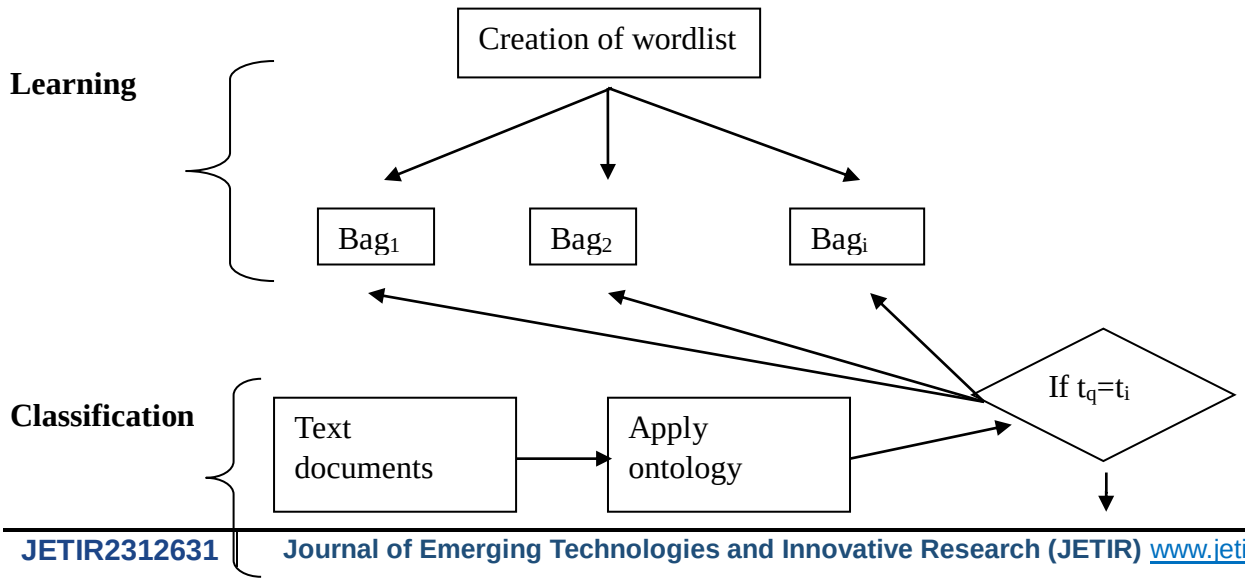
In this method, text classification based on support vector machine and description on Ontology representation is given. Support vector machine creates a hyper plane between the documents; the maximum distance between the hyper plane to document is taken and assign the document to that class. The process of classifying an unknown text document by the use of ontology representation is called as class wise ontology representation through the Bag of Words model.

3.3.5.1. Proposed model

The proposed model consists of two stages:
(a) Creation of class dependent term document matrix using wordlist
(b) Classification

3.3.5.2. Creation of Class Dependent Term Document Matrix

In this section, the term document matrix for each class is created. Given corpus C where C=[C_1, C_2, C_3 … , C_k], with n_i number of documents present in each class C_i
For each document $d_j \in C_i$ apply pre-processing techniques and create wordlist for the pre-processing document. Then create a Bag of Words (BOW) representation for wordlist. This procedure is carried out on all classes and thus the class wise Ontology representation is created through the bag of words model are created. Once the class wise ontology representation of words created, a term document matrix FM_i for i^{th} class is obtained, where $FM_i(l,j)$ denotes the frequency of occurrences of the term/word W_1 occurs in the j^{th} document of the i^{th} classs.



Assign class

Fig 3.3.5.2. Architecture of proposed model

Algorithm: Class dependent BOW creation

Input: Labelled documents of k classes
Output: $WG_i \forall i = 1,2,3,\ldots k$ classwise BOW

Method:
For each class C_i do
(Let there be n_i number of labeled documents in C_i)
$WG_i = \{\}$
For each document d_j in C_i do
Apply pre-processing
Tokenize
Stopword removal
Stemming
Creation of wordlist
WG_i = All the remaining words of dj
$WG_i = Union(WG_i, W_{ji})$
For end
Algorithm ends

3.3.5.3. Classification

Given an unlabelled document d_t, the labels of all classes for which it belongs are supposed to be predicted and it is subjected to the pre-processing steps to obtain its Wordlist and create a bag of words model for wordlist. Once the bag of words model is computed for wordlist , a feature vector F_t is obtained by the use of frequency of occurrences of each word in d_t with respect to a class C_i using WG_i, the wordlist of the class C_i. This feature vector F_t computed with respect to the class C_i, is then matched against each row of FM_i, the term document matrix of the class C_i to compute the maximum distance between the hyper plane to documents d_t of the class C_i. Hence due to n_i number of documents present in C_i, we end up with n_i number of distance comparison for d_t with respect to the class C_i. After computing distance each document create another FM_i with this FM_i, pooled documents d_t compared, if the comparison is matched any of the class in the FM_i then that document d_t is assign to that class. In this way the unlabelled document are classified using Ontology based on SVM classifier.

Algorithm: Labelling of unknown document

Input: **(1)** unlabeled document d_t
(2): $FM_i \forall i = 1,2,..k$ Term document matrix of all k classes
Output: labels of unlabeled document d_t

Method:
For a given unknown document d_t do
Apply pre-processing
tokenization
Stopword removal
Stemming
Creation of wordlist for all documents
W_t = All the remaining words of d_t

For each class C_i do
For each class C_i do
For each word W_i in WG_i do

3.3.5.4. Term Sequence

An important limitation of the VSM/BOW based model is its assumption of non-existent correlation between adjacent words and the loss of information in the representation. It maps each document of an arbitrary length to a vector of a determined dimension in which the vicinity of two terms depends on the identity numbers (ID's) they have in the vocabulary. The stream of the words is not lost and the documents are simply mapped by the

vocabulary in a sequence of ID's. This representation keeps the sequential information of the words that is w_t is the t^{th} word in the document but not the t^{th} entry in the vocabulary. This representation is not well used in the research since the methods working on sequences are more difficult to design. Usually it is used in text segmentation problems where the goal of the task is to find the topic boundaries inside a document proposed a term sequence classification method based on adaptive Markov model.

3.3.6. Phrase Based Representation

Words alone cannot represent the meaning; Phrase of words can give more knowledge compared to unit word. For example, the phrase machine learning has a highly specific meaning that is separated and distinct from the words as a bag of phrased rather than bag of words. The main problem is to convert a document into a bag of phrases, which potentially increases the number of features. If all the two words phrases in a document are used as features, then there will be same number of phrase instances as word instances, but the number of distinct features will potentially grow from t to t^2.

3.3.7. Spectral Clustering Based Representation

The spectral clustering usually clusters the data points using the top eigenvectors of graph Laplacian, which is defined on the affinity matrix of data points. From the graph partitioning perspective, the spectral clustering tries to find the best cut of the graph so that predefined criterion function can be optimized. From the perspective of dimensionality reduction, spectral clustering embeds the data points into a low dimensional space where the traditional clustering algorithm(k-means) is applied. One major drawback of these spectral clustering algorithms use the non-linear embedding which is defined only on training data. Embedding is computationally expensive for large data set.

3.3.8. Graph Based Representation

One of the most popular graph representation used to cluster the text document is universal networking language (ULN). The ULN represents meaning sentence by sentence information on a sentences is represented as a hyper-graph having concepts as nodes and relations as arcs. The hyper-graph is also represented as a set of directed binary relations. Concepts are represented as character-string called universal words (UW). The UW can be annotated with attributes which provide further information about how a concept is being used in the specific sentences.

3.4. Text Classification:

The problem of text classification is basically a task of partitioning the features space into regions, where each region representing a category. It is the establishment of correspondence between a document and a category to which it belongs,. Ideally, one would like to arrange this partitioning such that none of the decisions ever go wrong. When this cannot be done, one would like to minimize the probability of errors, or if some errors are more costlier than others, then the average cost of errors are considered.

3.5. Experimental result and Evaluation

After text classification, we need to measure the efficiency of the classification technique. In this state accuracy, recall, and precision are computed to measure the performance of the classifier. Each classifier is tested using testing dataset, and the classification is counted in four categories: true positives, true negatives, false positives and false negatives. In this context, positive means a document is classified as a member of the target class and negative means a document is classified as a member of the target class.

Let **TP, TN, FP and FN** respectively denote the number of the true positives, true negatives, false positives and false negatives. Let N be the total number of classification,

i.e., **N=TP+TN+FP+FN** (1.1)

Accuracy is simply defined as the number of correct classification divided by the total number of classifications.

i.e.., **Accuracy *(A)* =TP+TN/N** (1.2)

Precision measure the proportion of documents classified as positive that are correctly classified.

i.e., **Precision (*P*) =TP/TP+FP** (1.3)

Recall measures the proportion of documents which really are positive that are classified correctly.

i.e.. **Recall(*R*) =TP/TP+FN** (1.4)

In practice, increase in *P* is often gained at the expense of *R* and vice versa. An extremely conservative classifier will produce *P>>R*. This creates problems in comparing classification methods. In order to balance precision and recall, high-order statistics such as F-measure is suggested and it is defined as

i.e.., F-measure = 2*(*PR)/P+R*

.

3.5.1. DATA SET

Any system has to be empirically tested for its various characteristics to assess its effectiveness in meeting for which it has been designed. Many standard benchmark collections are available which can be used as initial corpora for text classification. In this work, we have used 20Newsgoup large dataset, 20Newsgoup Mini dataset and Reuters-21578 transcriptions standard datasets.

3.5.1.1. 20 Newsgroup Large Dataset

The 20Newsgroup Large (*http://people.csail.mit.edu/jrennie/20Newsgroups/)*dataset is a collection of approximately 20,000 newsgroup documents, partitioned nearly 20 different newsgroups. One thousand messages from each of the twenty newsgroups were choose at random and partitioned by the newsgroup name. Approximately 4% of the articles are cross posted. The articles are typical posting of other thus have headers including subject lines, signature files, and quoted portions of other articles. In this dataset, the data is organized into 20 different newsgroups, each corresponding to a different topic. The subject and body of each message from the content of a document, and the newsgroup names become the categories label. In addition, the category set has a hierarchal structure (“sci.space”, “sci.electronics”, “sci.med”, and “sci.crypt ” are subcategories of “sci(science)”).Table 3.5.1.1 shows the categories in 20 newsgroup Large and the indices of their respective documents.There is no fixed way to split 20Newsgroup into training set and a testing set.

Table 3.5.1.1: The categories of 20 Newsgroup Large Dataset

Category	Number of tests
alt.atheism	0-999
comp.graphics	1000-1999
comp.os.ms-windows.misc	2000-2999
comp.sys.ibm.pc.hardware	3000-3999
comp.sys.mac.hardware	4000-4999
comp.windows.x	5000-5999
misc.forsale	6000-6999
rec.autos	7000-7999
rec.motorcycles	8000-8999
rec.sport.baseball	9000-9999
rec.sport.hockey	10000-10999
sci.crypt	11000-11999
sci.electronics	12000-12999
sci.med	13000-13999
sci.space	14000-14999
soc.religion.christian	15000-15999
talk.politics.guns	16000-16999
talk.politics.mideast	17000-17999
talk.politics.misc	18000-18999
talk.religion.misc	19000-19999

3.5.1.2. 20 Newsgroup Mini Dataset

The 20Mini Newsgroup (*http://kdd.ics.ucl.edu/databases/20Newsgroups/20Newsgroup)*is a subset of the 20 newsgroup large dataset, which contains 100 randomly selected messages from each newsgroup. Large 20 newsgroup collection is used as a source to message from each newsgroup as it contains a range of topics that overlap to varying degrees. Create a mini newsgroup contains about 2000 documents evenly dived among 20 Usenet discussion groups. In 20mini newsgroup, there are 20 classes each with 100 documents which are randomly picked from the effectiveness of text classification algorithm that can run on a machine with lower processing speed. Table 3.5.1.2. shows the categories in 20 Newsgroup Large and the indices of their respective documents. There is no fixed way to split 20 Newsgroup into training set and a testing set.

Table 3.5.1.2.: The categories of 20 Newsgroup Mini Dataset.

Category	Number of tests
alt.atheism	0-99
comp.graphics	100-199
comp.os.ms-windows.misc	200-299
comp.sys.ibm.pc.hardware	300-399
comp.sys.mac.hardware	400-499
comp.windows.x	500-599
misc.forsale	600-699
rec.autos	700-799
rec.motorcycles	800-899
rec.sport.baseball	900-999
rec.sport.hockey	1000-1999
sci.crypt	1100-1199
sci.electronics	1200-1299
sci.med	1300-1399
sci.space	1400-1499
soc.religion.christian	1500-1599
talk.politics.guns	1600-1699
talk.politics.mideast	1700-1799
talk.politics.misc	1800-1899
talk.religion.misc	1900-1999

3.5.1.3. REUTERS - 21578 (Transcriptions)

REUTERS -21578 *(htpp://www.daviddlewise.com/resouse/testcollections/reuters21578/)* is one of the most widely used collection for text classification research, and it is the improved version of RCV1 (Reuters corpus volume 1). The data was originally collected and labelled by Carnegie group Inc. and Reuters Ltd., in the course of developing a text categorization system. Reuters (transcriptions), which are distributed in 10 files, mostly Concerning business and economics. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contains 200 documents, while the last file (reut2-000.sgm). The files in Reuters-21578 are in Standard Generalized MarkupLanguage (SMGL) format. Each of the 10 files begins with a document type declaration line: <!DOCTYPElewis SYSTEM "lewis.dtd">. Each Reuters tag contains explicit specifications of the values of five attributes, TOPICs, LEWISSPLIT, CGISPLIT, OLDID and NEWID. These attributes are meant to identify documents and group of documents. The following are the characteristics that this dataset possesses:

- It is multi-label. That is each document may belong to more than one category.
- The set of categories is not exhaustive. That is some documents belong to no category at all.
- The distribution of the documents across the categories is highly skewed, in the sense that some categories have very few documents classified under them while others have thousands.
- There are several semantic relations among the categories (there is a category WHEAT and a categories GRAIN, which are obviously related), but these relations are hidden (there is no explicit hierarchy defined on the categories).

Table 3.5.1.3 shows the categories in **REUTERS- 21578 (Transcriptions)** and the indices of their respective documents. There is no fixed way to split Reuters (transcriptions) into training set and a testing set.

Table 3.5.1.3. : The categories of REUTERS- 21578 (Transcriptions) Dataset.

Category	**Number of tests**
Earn	20
Acquisitions	20
Money-fx	19
Grain	20
Curd	20

Trade	21
Interest	20
Ship	20
Wheat	20
Corn	20

3.6. Result Table

Over view of Experimentation

In this section, we present the result of the experiments conducted to demonstrate the effectiveness of the proposed method on all three datasets viz.., 20Newsgroup large, 20News group Mini, Reuters.
During experimentation, we conducted six different sets of experiments. In the first set of experiments, we used 50% of the documents of each class of a dataset to create class representative vectors (training) and the remaining 50% of the documents for testing purpose. On the other hand, in the second set of experiments, the number of training and testing documents is in the ratio 60:40. In the third set of experiments, the number of training and testing documents is in the ratio 70:30.In the fourth set of experiments, the number of training and testing documents is in the ratio 80:20.In the fifth set of experiments, the number of training and testing documents is in the ratio 30:70.In the sixth set of experiments, the number of training and testing documents is in the ratio 40:60. All experiments are repeated 5 times by choosing the training samples randomly. As measures of goodness of the proposed method, we computed accuracy, precision, recall and F-measure. The maximum values of the classification of accuracy, precision, recall and F-measure are tabulated based on datasets and proposed methods.

3.6.1. Reuters (transcription) Dataset

In Reuters multi label dataset consist 10 classes, 200 documents represented in each class 20 documents are there. All proposed methods tabulated here of 5 trails.

Table 3.6.1.1 Result analysis of Reuters dataset of ratio 50%-50%

Trails	Accuracy	Precision	Recall	F-measure
1	98.7500	100	97.7778	98.8764
2	97.5000	95.2381	100	97.5610
3	88.7500	85	91.8919	88.3317
4	97.5000	97.3684	97.3684	97.3684
5	72.5000	71.1111	78.0488	74.4186

Table 3.6.1.2 Result analysis of Reuters dataset of ratio 60-40

Trails	Accuracy	Precision	Recall	F-measure
1	83	83.6735	82	82.8283
2	90	92.4528	89.0909	90.7407
3	94	88.2353	100	93.7500
4	87	85.1852	90.1961	87.6190
5	97	94.4444	100	97.1429

Table 3.6.1.3 Result analysis of Reuters dataset of ratio 70-30

Trails	Accuracy	Precision	Recall	F-measure
1	83.2298	60	73.8333	75.3330
2	84.4750	12.5000	15.3846	13.7931
3	79.5031	11.1111	25	25.3846
4	86.3354	13.1121	10	20.0003
5	91.3043	55.1724	94.1176	69.5652

Table 3.6.1.4 Result analysis of Reuters dataset of ratio 80-20

Trails	Accuracy	Precision	Recall	F-measure
1	35	37.8378	82.3529	51.8519
2	92.5000	83.3333	100	90.9091
3	95	95	95	95
4	97	100	75	85
5	85	57.1429	100	75

Table 3.6.1.5 Result analysis of Reuters dataset of ratio 30-70

Trails	Accuracy	Precision	Recall	F-measure
1	84.3972	15.3846	15.3846	15.3846
2	97.8723	100	78.5714	88
3	87.2340	14.2857	7.6923	10
4	87.2340	33.3333	5.8824	10.0000
5	90.7801	37.5000	27.2727	31.5789

Table 3.6.1.6 Result analysis of Reuters dataset of ratio 40-60

Trails	Accuracy	Precision	Recall	F-measure
1	81.8192	7.6923	9.0909	8.3333
2	81.8192	8.3333	8.3333	8.3333
3	85.9504	25	15.3846	19.0476
4	86.7769	10	12.5000	11.1111
5	99.1736	100	92.3077	96.0000

3.6.2. 20 Newsgroup Mini Dataset

In 20 Newsgroup Mini dataset consist 20 classes, 2000 documents represented in each class 100 documents are there. All proposed methods tabulated here maximum of 5 trails.

Table 3.6.2.1. Result analysis of 20 Newsgroup Mini Dataset of ratio 50-50

Trails	Accuracy	Precision	Recall	F-measure
1	93.5000	63.9344	47.5610	54.5455
2	96	70.7865	81.8182	75.9636
3	96	75	75	75
4	91	49.2537	37.0787	42.3077
5	91.4000	46.3415	22.8916	30.6452

Table 3.6.2.2. Result analysis of 20 Newsgroup Mini Dataset of ratio 60-40

Trails	Accuracy	Precision	Recall	F-measure
1	95	79.6296	59.7222	68.2540
2	95.3750	76.2712	66.1765	70.8661
3	94	66.6667	56.7164	61.2903
4	92.8750	64.5161	30.3030	41.2371
5	94.1250	62.7907	46.5517	53.4653

Table 3.6.2.3. Result analysis of 20 Newsgroup Mini Dataset of ratio 70-30

Trails	Accuracy	Precision	Recall	F-measure
1	93.8333	82.7586	42.8571	56.4706
2	96.6667	77.7778	77.7778	77.7778
3	94.6667	69.0909	71.6981	70.3704
4	95.5000	78	70.9091	74.2857
5	94	70.4545	57.4074	63.2653

Table 3.6.2.4. Result analysis of 20 Newsgroup Mini Dataset of ratio 80-20

Trails	Accuracy	Precision	Recall	F-measure
1	94.5000	73.5294	65.7895	69.4444
2	93.2500	68	47.2222	55.7377
3	95.2500	68.5714	75	71.6458
4	94.2500	71.0526	69.2308	70.1299
5	97	81.3953	89.7436	85.3659

Table 3.6.2.5. Result analysis of 20 Newsgroup Mini Dataset of ratio 30-70

Trails	Accuracy	Precision	Recall	F-measure
1	91.1429	44.1441	44.1441	44.1441
2	92. 1429	52	45.6140	48.5981
3	93. 1429	58.5586	56.5217	57.5221
4	92.6429	57.8947	46.6102	51.6432
5	96.5000	73.1343	88.2883	80

Table 3.6.2.6. Result analysis of 20 Newsgroup Mini Dataset of ratio 40-60

Trails	Accuracy	Precision	Recall	F-measure
1	100	100	100	100
2	95	66.9643	76.5306	71.4286
3	95	69.9029	75.7895	72.7273
4	95	67.5676	75.7576	71.4286
5	92.8333	60.4938	47.5728	53.2609

3.6.3 Comparative analysis

In this section, we compare the result of the proposed methods of SVM for different ratios for training and testing with different data sets.

Table 3.6.3.1.: Tabulates the classification Accuracy and F-Measure values of the SVM proposed methods

Training-Testing	DB1(A) SVM	DB2(A) SVM	DB1(F-M) SVM	DB2(F-M) SVM
30-70	98.7500	96	98.8764	75
40-60	97	95.3750	97.1429	70.8661
50-50	83	96.6667	75.3330	77.7778
60-40	97	97	85	85
70-30	97.8723	96	88	80
80-20	99.1736	95	96.0000	72.7283

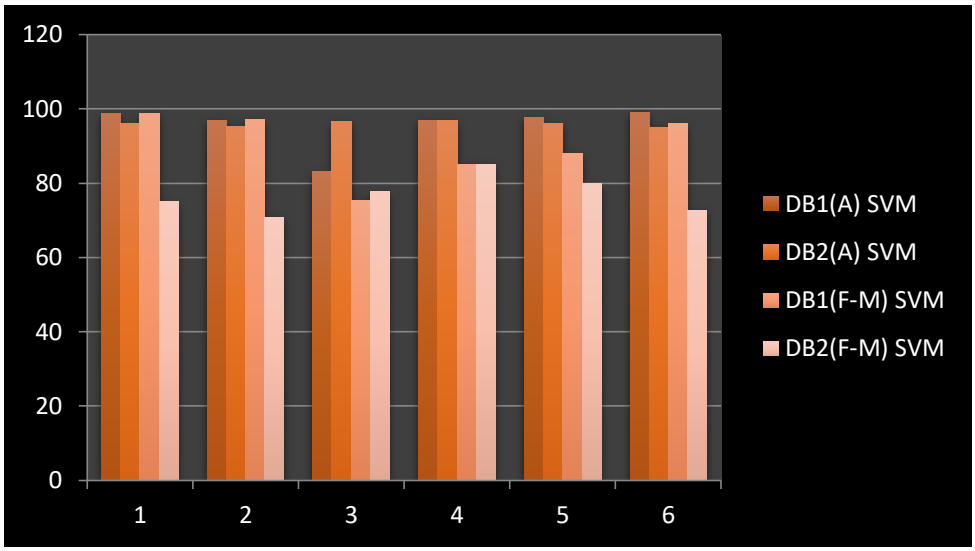


Fig 3.6.3.1 Comparative result analysis of Reuters and 20 Mini Newsgroup dataset

Note : DB1 – Reuters Data set, DB2 – Mini 20 Newsgroup Dataset

4. Conclusion

In this study, we delved into the realm of text categorization, leveraging the amalgamation of ontology and Support Vector Machine (SVM) classifiers to automate the classification of text documents. Our investigation encompassed the examination of datasets such as the 20 Newsgroup Large dataset, 20 Newsgroup Mini dataset, and REUTERS-21578, highlighting the significance of categorizing documents within predefined classes. Text categorization, a cornerstone of Natural Language Processing (NLP), assumes paramount importance in today's data-driven landscape. The exponential growth of unstructured data demands efficient methods to sift through and organize information. Our exploration elucidated the pivotal role played by automated classification in mitigating the challenges posed by voluminous unstructured data.

The utilization of ontology in tandem with SVM classifiers showcased promising outcomes, demonstrating the efficacy of knowledge representation and feature extraction in enhancing text categorization accuracy. The ability to represent relationships and terms within an ontology provided a structured framework for effective categorization, significantly streamlining the classification process. Moreover, our study shed light on the nuances between single-label and multi-label text classification problems, recognizing the significance of addressing documents that belong to multiple classes within a corpus. As we conclude, our findings underscore the pivotal role of automated text classification in various domains, ranging from information retrieval to decision-making processes. However, while our study marks a significant stride, there exist avenues for further exploration. Future research endeavours could delve deeper into refining the ontology-based approach, exploring diverse machine learning techniques, and extending the application of text categorization across domains beyond those studied in this work. In essence, the fusion of ontology and SVM classifiers offers a promising avenue for automated text categorization, unveiling opportunities for enhanced knowledge extraction and informed decision-making in an era inundated with unstructured textual data.

5. Future Work

- Extending the model for large dataset.
- Studying under different classifier other than SVM .
- Solving the imbalanceness problem of the corpus through adaptive boosting.

6. References

1. An Ontology- Based Framework for Text Mining. S. Bloehdorn and P. Cimiano and A. Hotho and S.Staab Institute AIFB University of Karlsruhe ,University of Kassel.
2. Classification and Assignment of Research Papers using Ontology based Hybrid Approach and RatishSrivastava and A.B.Bagwan, Department of Computer Engineering, JSPM's, RajarshiShahu College of Engineering, Pune University, Pune, India.

3. Ontological research paper selection using text mining and Kunj Patel, Dhananjay Rajput, Vasudeomadane,MayurShendge, A.E Patil.
4. Ontology Based Text Categorization - Telugu Documents and Mrs.A.KanakaDurga, Dr.A.Govardhan.
5. Text categorization using topic model and ontology networks and Yinghao Huang, Xipeng Wang and Yi Lu Murphey, Member, IEEE.
6. An Exercise in Ontology Driven Trajectory Simulation with Matalab Simulink and UmutDurak, SerdarGuler, S.KemalIder.
7. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. Westergaard, D., Stærfeldt, H. H., Tønsberg, C., Jensen, L. J., Brunak, S., &Rzhetsky, A. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. P L o S Computational Biology (Online), 14(2), [e1005962]. DOI: 10.1371/journal.pcbi.1005962
8. Exploration of Various Clustering Algorithms for Text Mining. a I.J. Education and Management Engineering, 2018, 4, 10-18 Published Online July 2018 in MECS (http://www.mecs-press.net) DOI: 10.5815/ijeme.2018.04.04
9. Computing and Information Systems is published normally three times per year, in February, May, and October, by the University of the West of Scotland.
10. A KNN Research Paper Classification Method Based on Shared Nearest Neighbor Yun-lei Cai, Duo Ji ,Dong-fengCai Natural Language Processing Research Laboratory, Shenyang Institute of Aeronautical Engineering, Shenyang, China, 110034
11. Text mining Ian H. Witten Computer Science, University of Waikato, Hamilton, New Zealand .
12. DaniYogatama and Noah A. Smith, (2014), Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers, Proceedings of the 31st International Conference on Machine Learning, Beijing, China, JMLR: W&CP.
13. Jialu Liu, Image Retrieval based on Bag-of-Words model, jialu.cs.illinois.edu/technical_notes/CBIR_BoW.
14. "Survey Paper on Feature Extraction Methods in Text Categorization" International Journal of Computer Applications (0975 – 8887) Volume 166 – No.11, May 2017
15. Hybrid feature selection methods for high-dimensional multi-class datasets Amit Kumar Saxena; Vimal Kumar Dubey; John Wang Department of Computer Science and Information Technology, Guru GhasidasVishwavidyalaya, Bilaspur, 495009, India ' Department of Computer Science and Information Technology, Guru GhasidasVishwavidyalaya, Bilaspur, 495009, India ' Department of Information Management and Business Analysis, Montclair State University, Montclair, NJ 07043, USA.
16. Filter Methods for Feature Selection – A Comparative Study, Noelia S´anchez-Maro˜no, Amparo Alonso-Betanzos, and Mar´ıaTombilla-Sanrom´an University of A Coru˜na, Department of Computer Science, 15071 A Coru˜na, Spain nsanchez@udc.es, ciamparo@udc.es, infmts00@ucv.udc.es
17. ARABIC TEXT CLASSIFICATION USING NEW STEMMER FOR FEATURE SELECTION AND DECISIONTREES,SAID BAHASSINE1, ABDELLAH MADANI2, MOHAMED KISSI3,1LIMA Laboratory, Department of Computer Science, ChouaibDoukkali University, Faculty of Science, B.P. 20, 24000, El Jadida, Morocco 2LAROSERI Laboratory, Department of Computer Science, ChouaibDoukkali University, Faculty of Science, B.P. 20, 24000, El Jadida, Morocco 3LIM Laboratory, Department of Computer Science, HASSAN II University Casablanca, Faculty of Sciences and Technologies, B.P. 146, 20650, Mohammedia, Morocco.
18. Feature Extraction and Duplicate Detection for Text Mining: A Survey, By Ramya R S, Venugopal K R, Iyengar S S& Patnaik L.
19. Survey Paper on Feature Extraction Methods in Text Categorization, DixaSaxena Department of Computer Science & Engineering MANIT, Bhopal ,S. K. Saritha, PhD Department of Computer Science & Engineering MANIT, Bhopal ,K. N. S. S. V. Prasad Department of Computer Science & Engineering MANIT, Bhopal
20. Global Journal of Pure and Applied Mathematics. ISSN 0973-1768 Volume 13, Number 7 (2017), pp. 3249-3260 © Research India Publications http://www.ripublication.com
    Comparative Analysis between Rough set theory and Data mining algorithms on their prediction, M. Sudha Assistant Professor, Department of Mathematics. Amet University,Kanathur,Chennai-600112, India. A. Kumaravel Dean, School of Computing, Bharath University, Selaiyur, Chennai-600073, India.