International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

An Automatic Text Document Classification using Modified Weight and Semantic Method

K.Meena, R.Lawrance

Abstract: ***Text mining is the process of transformation of useful information from the structured or unstructured sources. In text mining, feature extraction is one of the vital parts. This paper analyses some of the feature extraction methods and proposed the enhanced method for feature extraction. Term Frequency-Inverse Document Frequency(TF-IDF) method only assigned weight to the term based on the occurrence of the term. Now, it is enlarged to increases the weight of the most important words and decreases the weight of the less important words. This enlarged method is called as M-TF-IDF. This method does not consider the semantic similarity between the terms. Hence, Latent Semantic Analysis(LSA) method is used for feature extraction and dimensionality reduction. To analyze the performance of the proposed feature extraction methods, two benchmark datasets like Reuter-21578-R8 and 20 news group and two real time datasets like descriptive type answer dataset and crime news dataset are used. This paper used this proposed method for descriptive type answer evaluation. Manual evaluation of descriptive type paper may lead to discrepancy in the mark. It is eliminated by using this type of evaluation. The proposed method has been tested with answers written by learners of our department. It allows more accurate assessment and more effective evaluation of the learning process. This method has a lot of benefits such as reduced time and effort, efficient use of resources, reduced burden on the faculty and increased reliability of results. This proposed method also used to analyze the documents which contain the details about in and around Madurai city. Madurai is a sensitive place in the southern area of Tamilnadu in India. It has been collected from the Hindu archives. This news document has been classified like crime or not. It is also used to check in which month most crime rate occurs. This analysis used to reduce the crime rate in future. The classification algorithm Support Vector Machine(SVM) used to classify the dataset. The experimental analysis and results show that the performances of the proposed feature extraction methods are outperforming the existing feature extraction methods.***

Keywords: crime news, descriptive type answers, feature extraction, semantic similarity and text document classification.

I. INTRODUCTION

Text mining is the procedure used to derive useful information from the text. The various text mining tasks are text classification, document summarization, sentiment analysis and text clustering.

Text analytics involves information transformation, pattern recognition, information retrieval, association analysis, predictive analytics and visualization. The goal of text mining is to turn text data into useful data for analysis with the help of analytical methods and Natural Language Processing (NLP). The increased use of text management in the internet has resulted in the enhanced research activities in text mining.

Now-a-days, everyone has been using internet for surfing, posting, creating blogs and for various purpose. Internet has massive collection of many text contents. The unstructured text content size may be varying. For analyzing the text contents of internet efficiently and effectively, it must be classified or clustered. The collections of text document in the internet have been separated and stored each collection as dataset based on that need. The increased use of text management in the internet has resulted in the enhanced research activities in text mining.

Several preprocessing techniques, clustering and classification techniques are available for processing the text document. But the efficiency and effectiveness of the algorithms are not very significant. The consistency and certainty of the existing methods are not noteworthy for the text document classification and clustering. In the classification process, selection of preprocessing, feature extraction methods and classification methods are important for assigning class label to the text document.

Text mining processing steps are text document collection, preprocessing, feature extraction, incorporation of data mining techniques, interpretation and evaluation. The text documents are preprocessed using pruning and stemming process. Pruning means removing common words such as an, the, at, for, etc. After performing pruning, the words are given as input to the stemming procedure. This procedure produces the stem of the given word. These processes are used to reduce the size of the document. The important features from the text have been selected by using statistical models. Several feature extraction method discussed by various researchers.

Text representation is the important problem in information retrieval, natural language processing and text mining. It is used to represent the unstructured text as a numeric value. This format is more suitable for processing them computationally. Researchers used various techniques for text representation to reduce the gap between the semantic and syntactic relationship between words in the text document.

Uysal et al. [1] illustrate how much preprocessing is important for text classification. This analysis is takes place with different aspects such as text domain, dimension reduction, text language and classification accuracy. E-mail and news of two different languages such as English and Turkish are considered for analysis.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
K.Meena*, Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India. Email: msdmeena@gmail.com
R.Lawrance, Director, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamilnadu, India, Email:lawrancer@yahoo.com

The feature extraction used vector space models to the text document into vectors. The feature selection used methods such as information gain, gini index, document frequency and chi-square for select the important features from the text document. Finally, the classification steps used algorithms such as naïve bayes, decision trees, artificial neural network and support vector machines. Experimental results shows that exact mixture of preprocessing task provide vital upgrading of classification accuracy. Bullinaria et al. [2] proposed best computational method to take out more meaningful information from the large document. This method uses function word stop list, stemming and Single Value Decomposition (SVD) method for dimensionality reduction. The most common words in the document are called as functional words. If these functional words are removed from the document, it will reduce the processing time. SVD is used for dimensionality reduction and it also improves the performance of document analysis. Moh'd A Mesleh, Abdelwadood [3] proposed a procedure for Arabic text classification based on SVM. This procedure use chi square method for feature selection. Text classification classify the documents depend on the content of the document. Feature selection methods chi-squared statistics, term strengths, frequency thresholding, mutual information and information gain were compared with each other with Arabic data set. After comparison of various feature selection methods, chi-square method is best for Arabic data set. Chi square method gives better results compared with other method for text classification. Various classification methods such as KNN classifier, naïve bayes classifier and SVM classifier are considered for Arabic data set classification. Comparison result shows that SVM linear gives more appropriate result when compared with other classifiers. Zareapoor et al. [4] compare dimension lessening techniques for text classification. This paper takes email text for classification. Dimensionality reduction in the text data set can be done with the help of feature selection or feature extraction methods. This paper discusses and compares two dimensionality reduction methods and selects the best method for that purpose. Feature selection methods such as Information Gain Ratio (IGR) and chi square and feature extraction techniques such as Latent Semantic Analysis (LSA) and Principle Component Analysis(PCA) are considered for analysis. Feature extraction method is used to transform the data set into new set of features without removing them but create the new features set instead of the original dataset. Feature selection method is used to select the subset of features from original data set then it is considered for training and testing the classifiers. The result shows that the feature extraction methods give constant classification results. Zhang et al. [5] compares Term Frequency- Inverse Document Frequency (TF-IDF), multiword and Latent Semantic Indexing (LSI) for text representation. TF-IDF is used to select features from the text document. This method checks whether or not given term is related to the given document. If the term is related to the document means, consider that term is the important one and also assign the weight to the term. The drawback of the TF-IDF method is in dimensionality. The size of the feature set return by the TF-IDF is equal to the vocabulary size of the original document. It automatically increases the computation time. LSI produces the low dimensional structure text representation of the document. This structure represents the relationship between the terms. This method uses Single Value Decomposition (SVD) to reduce the size of the document. This paper compares information retrieval and text classification performance with these three feature selection method. This comparison shows that the multi word and LSI performs well than the TF-IDF. The performance of the classification mainly depends on the number of dimension of the text matrix. Jivani, Anjali Ganesh. [6] discussed various stemming algorithms with their advantages and disadvantages. Porter's developed a framework for stemming[7]. This is called as 'Snowball'. This algorithm is used for stemming in this proposed work.

This proposed work analyses some of the feature extraction methods and propose the enhanced methods for feature extraction. The existing feature extraction method Term-Frequency-Inverse Document Frequency(TF-IDF) assigned weight to the term based on the occurrence. But the proposed modified TF-IDF(M-TF-IDF) assigned weight to the term based on the occurrence and importance of the terms in the document. This weighting scheme used to increase the accuracy of the classification. But this method does not consider semantic similarity of the term. Hence Latent Semantic Analysis(LSA) method is discussed to select the terms based on the semantic similarity. The combination of M-TF-IDF and LSA assigned weight to the terms based on the importance and semantic similarity between the terms.

II. NEED FOR FEATURE EXTRACTION

Feature extraction is a technique used to extract the most relevant features for processing. This technique is related to dimensionality reduction. This reduced text document enhances the classification accuracy. This process is important in information retrieval and text mining. Feature extraction is an important task in text classification. It directly affects the classification accuracy. Vector space model is a base for feature extraction process. This model represents the text in an n-dimensional space. Each element in the n-dimension represents the features of the text document.

Feature extraction is useful to increase the efficiency, accuracy and scalability. It is a used for dimensionality reduction. If the size of the document is large, the performance of the classification algorithm is less. It is helpful as it can remove the curse of dimensionality or unrelated features from the text document. This is the first stage of knowledge discovery.

Extracted features from text used for analyzing different applications, automated terminology management, information retrieval, web mining, clinical records and research subject identification etc.

Text represented using vectors base on the previously discussed methods. The size of the vectors may be in different size based on the documents size. If the document size is too large, dimension of the vector is also high. High dimensional vector space takes too much time for classification. To solve this time complexity problem, dimensionality reduction method is used. To reduce the size of the features, some of the dimensionality reduction method such as Principal Component Analysis(PCA), Linear Discriminant Analysis(LDA) and Non-negative matrix factorization(NMF).

III. FEATURE EXTRACTION METHODS

A. N-gram

N-gram is a series of N words. This model is a probabilistic model used for finding the next item in the sequence. It is used in computational linguistics, data compression, communication theory and computational biology. This method is more useful because of its simplicity and scalability.

Two-word sequence of words is called as 2-gram or bigram, three-word sequence of words are called as 3-gram or trigram. Example for 2-gram,

Before you judge, understand why.

2-gram tokens are

{“before you”, “you judge”, ”judge understand”, “understand why”}

3-gram tokens are

{“before you judge”, ”you judge understand”, “judge understand why”}

The probability of a series of words is calculated depend on the multiplication of probabilities of each word.

N-gram method is used for automatic spelling correction, automatically generates text from speech and also used to determine the relationship between the words in the text.

Major drawback of this method is sparcity. Sparse data for low frequency affect the classification performance. This model does not capture the long range dependencies.

B. Bag-of-words

This method is flexible and simple which is used for extracting features from the text. It is an algorithm used in information retrieval, document classification and natural language processing. In this model, a document or a sentence is considered as a bag holding words. This method analyses and identifies the different bag of words and classifies the bag of words by matching. This method first constructs the vocabulary from the given document. It consists of key and value. Key indicates the term in the document and the value indicates the number of times the word occurs within the document.

The following example shows how BOW works.

Consider the 2 text documents

“rose is a beautiful flower”

“rose is used for prepare perfume”

BOW first form the dictionary

{ ’a’:1, ’beatuiful’:2, ’flower’:3, ’for’:4, ‘is’:5, ‘prepare’:6, ’perfume’:7, ‘rose’:8, ‘used’:9}

Vectors of BOW are

[1, 1, 1, 0, 1, 0, 0, 1, 0]

[0, 0, 0, 1, 1, 1, 1, 1, 1]

In the first document, rose occur 1 time, beautiful occur 1 time, 0 count for prepare and so on.

The disadvantages of BOW method is it used one hot encoded vector method, for example, the size of the vocabulary is n, each words in the document is represented by n dimensional vector space with 1 in the index corresponding to the word and 0 in the every other index. If the vocabulary size increases, the BOW model faces the scalability challenges. Sparse representation of this model occupies more space and also increases the computational time. Meaning of the word and context are discarded in this representation but both are very important in the classification process. Hence, this method is not suitable for feature extraction for text classification.

C. TF-IDF

This technique is used to select features from the text document. Term frequency represents the number of times the terms occur in the document. The weight of the term is the proportional to the term frequency. The TF gives weight to all of the words in the document without considering the significance of the terms. The Inverse Document Frequency(IDF) is combined with TF to give more weight to the rarely occurring terms and reduce the weight of the frequently occurring terms. If a term does not occur in the specific document or if it appears in every document, this method assign zero to that term. Jing et al. [8] applied TF-IDF for text classification. To classify the text data effectively, choosing the significant features using TF-IDF from the data set is essential. This calculation used to find out the significant words in the text document and also tells us about what this document talks about.

TF is calculated by dividing the count of number of times the word ‘t’ appears in the document by counting the total number of words in the document. It is shown in equation (1).

For example, consider a data set with many documents. Let us consider D1, D2, D3 …. .. Dn is a set of documents and its total collection is named as ND.

TF = (Tt/NTD) (1)

where Tt – the word t repeated how many times in the text document

NTD – count of words in the text document

The IDF is calculated by the total number of documents divided by the number of documents containing the term, and then taking the logarithm of that quotient as given in equation in(2).

$$IDF = \log(\mathrm{ND}/(1 + \mathrm{NDt}) \quad (2)$$

where ND – total number of documents

NDt – count of how many documents containing the word t

TF-IDF assigns weight to the terms rarely occurring in the document and it simplifies the computation process. This method improves the precision and recall and overall it increases the performance of the classification. The main drawback of this method is it does not consider the semantic similarity between the words because the words are represented as an index.

D. Linear Discriminant Analysis

Linear discriminant analysis is mainly used for dimensionality reduction method. LDA is used in pattern recognition, statistics and machine learning to determine the features that separate the different classes of objects. This method is similar to regression analysis and analysis of variance method. In this method, one variable is expressed as combination of other measurements or features. LDA uses continuous dependent variable and categorical independent variable.

This method is used when categories are identified priori. The discriminant analysis is a technique used to analyze the data when the dependent variable is categorical or the independent variable is interval in nature. It is similar to regression analysis and has different benefits like statistical tool. This method is used to find which class variable is related to the dependent variable. PCA and LDA both are linear transformation techniques but LDA is a supervised process and PCA is an unsupervised process. LDA tries to find out the features subspace which maximizes the class separability. PCA finds out the directions of maximal difference.

Linear discriminant analysis has been depends on the model of searching which combination of variables separate the classes.

The linear discriminant analysis finds out the label of new point if the points satisfy the following condition in (3)

$$\beta^{T}\left[P-\left[\frac{M_1}{M_2}\right]\right] > -\log\left[\frac{P(co_1)}{P(co_2)}\right] \qquad (3)$$

where β^T represents the vector of coefficients, P represents the vector of data, M_1 and M_2 represents the mean vector, CO_1 and CO_2 represents the matrices of covariance and $P(CO_1)$ and $P(CO_2)$ represents the probability of class.

The limitations of LDA are it is parametric method because LDA not preserve the complex structure when the distributions are no-gaussian. This method produces only c-1 projections because some other method may be employed when the classification error estimates that some extra features are needed. When the discriminatory function is in the variance of data not in the mean, this method will fail.

There are different types of LDA are available such as orthonormal LDA, non-parametric LDA, generalized LDA and multilayer perceptrons.

E. Principal Component Analysis(PCA)

Principal Component Analysis is also one of the dimensionality reduction methods which are used to decrease the dimensionality of large data sets. It reduces the dimensionality but it contains same information like large data sets. If the large data set is reduced into smaller data set, it is very easy to visualize and explore and also used to increase the speed of the machine learning algorithms. It is a method used for extracting vital variable from a large set of variables from the data set. It is used to capture essential information and also used to reduce the size of the data set. PCA contains the biggest amount of the variance of the original data set. It produces a new subset of attributes from the original attributes. This technique is useful if the dataset contains more number of independent variables but they contain high correlation between them. PCA is performed on a symmetric correlation matrix. It should be numeric and have standardized data. In PCA, first principle components are identified. A principle component is a standardized linear combination of the original predictors in a data set. Executing PCA on unnormalized data will lead to large loadings for variables with high variance. Before finding the principle component, data should be normalized. For given m*n dimensional data, min (m-1, n) principle components can be constructed where m represents the number of documents n represents the number of terms. The correlation between these components should be zero. It is an unsupervised approach because the response variable is not used to determine the direction of these components.

PCA first normalize the continuous initial variables. This method enables each variables contribute equally to the analysis. This can be done by using the formula (4)

$$Z = \frac{x-\mu}{\sigma} \qquad (4)$$

Where x represents the value, μ represents the mean value and σ represents the standard deviation.

Then calculate the covariance matrix to determine the relations between the variables. This matrix is a symmetric matrix has covariances connected with all possible pairs of variables. The covariances informed that the relations between the variables. If the covariance value is positive, the two variables decrease or increase together. If the covariance value is negative, one variable is inversely correlated with other variable.

Then to find out the principal component analysis, calculate the eigenvales and eigenvectors of the covariance matrix. The eigenvectors of the covariance matrix are principal components of the large data set. The eigenvectors are orthogonal because the covariance matrix is symmetric. Each eigenvector has a eigenvalue. It is a scalar value. If the eigenvalue is large, then the principal components describes that there exists a large amount of variance in the data. If the eigenvalue is small, principal components describes that there exists a small amount of variance in the data. If the eignevalue is zero, components describes that there exists no variance in the data.

The resultant principal components are new variables that are built from the initial variables. PCA deposit maximum amount of information in the first component, then remaining maximum amount of information in the second component and so on.

The limitations of PCA are the selection principal components are depending on the scaling of the variables, PCA only detain the linear correlations between the features and the mean removal process.

F. Non-negative matrix factorization

Non-negative matrix factorization is a collection of algorithms in linear algebra and multivariate. In this matrix, all elements are positive. It gives an unsupervised linear description of data. NMF is more useful if the input data have large number of attributes. It can produce needed output by combining the attributes. It is used to extract meaningful features from the high dimensional data and represent the features using non-negative matrix. NMF induces sparsity and leads to part based decompositions. This matrix is represented by the combination of two matrices.

Let matrix N be the multiplication of matrices P and Q. It is shown in the equation (5).

$$N = P*Q \qquad (5)$$

If N is an m*n matrix, P is an m*p matrix and Q is a p*n matrix. Here p is less value than both n and m. In this matrix, P represents the basic features of data points and Q tells us which feature is present in which data point. Different types of NMF are convex NMF, nonnegative rank factorization, online NMF and approximate NMF. NMF are used in face recognition, text classification, recommender systems, astronomy, spectral data analysis, bioinformatics and topic modeling.

In text mining, documents are represented as matrix. In this matrix, each row represents the word and each column represents the document. NMF has been produced two matrices P and Q. The documents are represented as P matrix and Q matrix tells that how to sum contributions from diverse topics to restructure the word mix of a given document.

In image processing, NMF used for sparse coding, feature representation and video tracking. In bioinformatics, NMF has been used for microarray data analysis. NMF also used for blind source separating, acoustic signal processing and so on.

Hyperspace Analog to Language

Hyperspace analog to language (HAL) is another model of semantic memory which relies on statistical correlations in the input text to extract semantic information. A brief description of HAL follows. The basic methodology of the simulation is to develop a matrix of word co-occurrence values for a given vocabulary. A "window" of a certain size (*e.g.*, ten words) is defined which is slid over the corpus. The co-occurrence values are inversely proportional to the number of words separating a specific pair of words. For example, a word pair separated by a nine-word gap would gain a co-occurrence strength of one, while the same pair appearing adjacent to one another would receive an increment of ten in the matrix. The window is moved over the entire text corpus, with the co-occurrence of the center word in the window scored with all the other words in the window. This produces an N-by-N matrix, where N is the number of words in the vocabulary. Each row in this vector represents the degree to which each word in the vocabulary preceded the word corresponding to the row, while each column represents the co-occurrence values for words following the word corresponding to the column. A full co-occurrence vector for a word consists of both the row and the column of that word.

When a human meets a new concept, derive its meaning via an accumulation of contexts. On the basis of distributional characteristics of semantics, different lexical semantic space models have been examined. In the language, the meaning of the word is taken by referring its co-occurrence pattern with the collection of text. Document spaces and corpus of text are the two major classes of semantic space models. The former represents words as vector spaces of text fragments (e.g. documents, paragraphs, etc.) in which they occur. HAL automatically constructs a dimensional model from a complete collection of text. HAL is an N*N matrix with an N-length vector for each unique word in the corpus. HAL needs to be searched the corpus word by word and for each word assigning a value to other words in its neighborhood. All words within the window are considered as co-occurring with each other with strengths inversely proportional to the distance between them. Thus, a corpus is converted to a high dimensional semantic space, with minimal consideration to grammar. Given an n-word vocabulary, HAL space is a word-byword matrix constructed by moving a window of length l over the corpus by one word increment, ignoring punctuation, sentence and paragraph boundaries [8]. The distance between two words within the window is d, then the weight of an association between them is computed by (l-d+1). HAL produced an accumulated co-occurrence matrix for all the words in a target vocabulary after traversing the whole corpus. HAL is direction sensitive. In the co-occurrence matrix, the row vectors represent the co-occurrence information for the words preceding every word and the column vectors represent the co-occurrence information for words following it. An illustration for the HAL space is depicted in the Table I using a 5 word window (l=5).

TABLE I: EXAMPLE OF A HAL SPACE

A) An example sentence: Computer Mouse Keyboard Laptop Computer Mouse

	Computer	Mouse	Keyboard	Laptop
Computer	0	0	0	0
Mouse	5	0	0	0
Keyboard	4	5	0	0
Laptop	3	4	5	0

B) Vector for "Keyboard" – 45000005

4.1.A) Example of HAL matrix of the sentence "Computer Mouse Keyboard Laptop". Window sizes 5, co-occurrences between a particular word and those that precede it are encoded in rows, while those that follow it are encoded in columns.

B) Example co-occurrence vector for the word "Keyboard" in the above matrix is 45000005. Column co-occurrence values appear after row co-occurrence values.

HAL does a good job of separating categories of words. HAL provides a very high-dimensional sparse matrix representation of the word co-occurrence in a text corpus.

G. Latent Semantic Analysis

Human beings understand a new concept by deriving its meaning through an accumulation of contexts in which the concept appears. [9] In natural language processing, various semantic models have been used. The two major types of semantic space models are document spaces and word spaces. The document space represents words as vector spaces of text fragments such as documents and paragraphs in which they occur. LSA is an example for document space representation. The word space represents words as vector spaces of other words which co-occur with target words within a certain distance. Hyperspace Analog to Language is an example for word representation. Vector space representation is used for the measurement of similarity between words.

Latent Semantic Analysis is a natural language processing technique used for analyzing the relationship between the set of documents and the terms. This technique uses no humanly constructed dictionaries, semantic networks, knowledge bases, syntactic parsers, grammars or morphologies. LSA literally means analyzing documents to find the underlying meaning or concepts of those documents. It is a fully automatic mathematical technique for extracting and inferring relations of expected contextual usage of words in the passages of discourse which returns a matrix. In this matrix, rows represent the unique words and columns represent each paragraph. This is a sparse matrix because most terms occur in few documents and it is a large matrix because there are many terms across many documents. Thus the matrix is reduced to discover its latent properties. Hence, it used a mathematical technique for takeout and deduces relations of words in the passages and returns a matrix. Danielle et al. [10] proposed a latent semantic analysis method for extracting and representing the meaning of words using statistical computations applied to large corpora of text.

Since the advent of LSA, researchers have developed and tested alternative statistical methods designed to detect and analyze meaning in text corpora. This research exemplifies how statistical models of semantics play an important role in our understanding of cognition and contribute to the field of cognitive science.

Latent Semantic Analysis (LSA) is a high-dimensional linear associative model that analyzes a large corpus of natural text and generates a representation that captures the similarity of words and text passages. A brief description of the LSA methodology follows. Input to LSA is a matrix consisting of rows representing unitary event types and columns representing contexts in which instances of the event types appear. One example is a matrix of unique word types by many individual paragraphs in which the words are encountered, where a cell contains the number of times that a particular word appears in a particular paragraph. Thus, this matrix represents raw, first-order co-occurrence relations between stimuli and the local contexts of episodes in which they occur. Then, each cell is transformed by taking the log of the frequency value in the cell. This approximates the standard empirical growth functions of simple learning. This also yields a kind of spacing effect; the association of A and B is greater if both appear in two different contexts than if they each appear twice in one context. In the second step, all cell entries for a given word are divided by the entropy of that word over all its contexts. This inverse entropy measure estimates the degree to which observing the occurrence of a component specifies what context it is in. The larger the entropy, the less information it provides about its context, and in turn, the less-usage defined meaning it acquires. The next step in LSA is the Singular Value Decomposition (SVD) of the transformed matrix. It is a general method for linear decomposition of a matrix into independent principal components. SVD converts the transformed associative data into a condensed representation, which captures the higher-order, indirect associations. For example, if a particular stimulus, X (*e.g,*. a word) has been associated with another stimulus, Y, by being frequently found in joint context, and Y is associated with Z, then condensation causes X and Z to have similar representations. The strength of the association XZ depends not only on the combination of the strengths XY and YZ, but also on relation of X, Y, and Z to every other entity in space.

The mathematical technique of Single Value Decomposition (SVD) is used to reduce the matrix. Latent Semantic Analysis constructs a rectangular matrix of words by passages, with each cell containing a transform of the number of times that a given word appears in a given passage. The matrix is then decomposed in such a way that every passage is represented as a vector whose value is the sum of vectors standing for its component words. Similarities between words and words, passages and words are then computed using cosine similarity function. LSA is one of the best dimensionality reduction method[11]. The advantages of LSA are it is easy to implement, understand and use its works well on dataset with diverse topics, this method is reliable and faster than other dimensionality reduction methods and LSA acted as a good predictor of information retrieval process.

LSA first construct the term document matrix(A) for the given document. Then the document to document matrix(B) is constructed by multiplying the transpose of A by A. The term to term matrix(C) is constructed by multiplying the A by the transpose of A. Both B matrix and C matrix are square and symmetric. Then calculate the SVD on A using the matrices B and C. A is decomposed into the combination of three matrices. It is shown in the equation (6).

$$A = U \sum V^T \quad (6)$$

Where U represents the matrix of eigenvectors of B, V represents the matrix of the eigenvectors of C and $\sum$ is the diagonal matrix of the singular values obtained as square roots of the eigen value of B.

The singular values along the diagonal of $\sum$ are listed in descending order of their magnitude. Some of the singular values are too small and it is negligible. Ignore these small singular values and replace them by zero. Therefore for calculation purposes, can only keep the n singular values in $\sum$. Then $\sum$ will be all zero except the first n entries along its diagonal. As such, can reduce the matrix $\sum$ into $\sum_n$ which is an n*n matrix, containing only the n singular values and also reduce the matrix U and V^T into U_n and V_n^T to have n columns and n rows respectively. Now matrix A is approximated and shown in equation (7).

$$A = U_n \sum V_n^T \quad (7)$$

This matrix A is the reduced matrix that is the dimension of the original matrix is reduced but it contain all the terms that are useful for further processing. The multiplication of $U \sum$ is used to measure the similarity between the two terms. The multiplication of $V^T \sum$ is used to measure the similarity between the two documents. To check whether the given term is within a document or not, multiply the $U \sum$ and $V^T \sum$ then divide it by 2. The documents are retrieved based on the similarity value returned by this multiplication. This process is used in the information retrieval applications.

This Feature extraction is used to convert the original matrix into the new with reduced size. This reduced matrix also consists of features same as original matrix but the size is reduced without delete the features selected. If the number of features in the input data is large, it will take more to process. Hence, to reduce the size of the matrix without reducing the original features, the feature extraction method is used. The extracted features are considered for further processing. The reduced matrix constructed after feature extraction process is used to increase the classification accuracy and also it take less time to classify the documents.

IV. PROPOSED FEATURE EXTRACTION METHOD

A.Modified TF-IDF(M-TF-IDF)

TF-IDF used to give weight if particular word occurs within the particular document or if it does not appears in every document. This formula is enhanced to increase the weight of the most important words and decreases the weight of the less important words. The modified weight is shown in equation (8).

Weight1=(Tt/NTD)*log(ND/(1+NDt))*(ND/NTD) (8)

Where NTD – count of words in the text document

ND – total number of documents

This weight formula first finds the number of times the word occurs within the document.

Then find the rareness of the word that is checked whether the word occurs in each document or the word occurs within the particular document. If the word occurs in each document, no need to assign weight to the word. If the word occurs in the particular document only, assign some weight to the word. Then find the importance of the word. The existing TF-IDF method assign only weight to the terms based on the occurrence. But the enlarged TF-IDF increase the weight of the important word in the document and the less important words weight is decreased. This method check the number of terms selected for further processing in the particular document. If the less number of terms are selected from the document, assign more weight to the terms. If the number of words selected from the document is more, assign less weight to the terms.

For example, consider five sample documents like

Doc 1="This is a sample"
Doc 2="This is another example"
Doc 3="Rose is a beautiful flower"
Doc 4="Most of the flowers has five petals"
Doc 5="Rose is used for prepare perfume"

Create the corpus for the documents. After clean the corpus, construct the document term matrix. Then the TF-IDF and M-TF-IDF methods are used to assign the weights to the selected terms. The M-TF-IDF has been assigned weights to the term based on the occurrence and number of words selected from the document. If less number of words selected from the document, M-TF-IDF assigned less weights to the term. If more number of words selected from the document, M-TF-IDF assigned more weights to the term. Here, discussed some of the words weight assigned by TF-IDF and M-TF-IDF. The following figure 1 and figure 2 shows the selected terms and their weights assigned by using TF-IDF and M-TF-IDF. The term 'beautiful' weight is increased after applied M-TF-IDF, because from this document only 3 words are selected. The term 'petals' weight is decreased after M-TF-IDF has been applied because from that document 6 words are considered. The term 'perfume' weight is not changed because from that document 5 words are considered.

TF-IDF("beautiful") = 0.7739
M-TF-IDF("beautiful")=0.7739*(5/3)=1.2898
- After applied M-TF-IDF, the weight of the term "beautiful" is increased.

TF-IDF("petals") = 0.39
M-TF-IDF("petals") = 0.3869*(5/6) = 0.3248
- After applied M-TF-IDF, the weight of the term "petals" is decreased.

TF-IDF("perfume") = 0.4644
M-TF-IDF("perfume")= 0.4644 * (5/5) = 0.4644
- After applied M-TF-IDF, there is no change in the weight of the term "perfume".

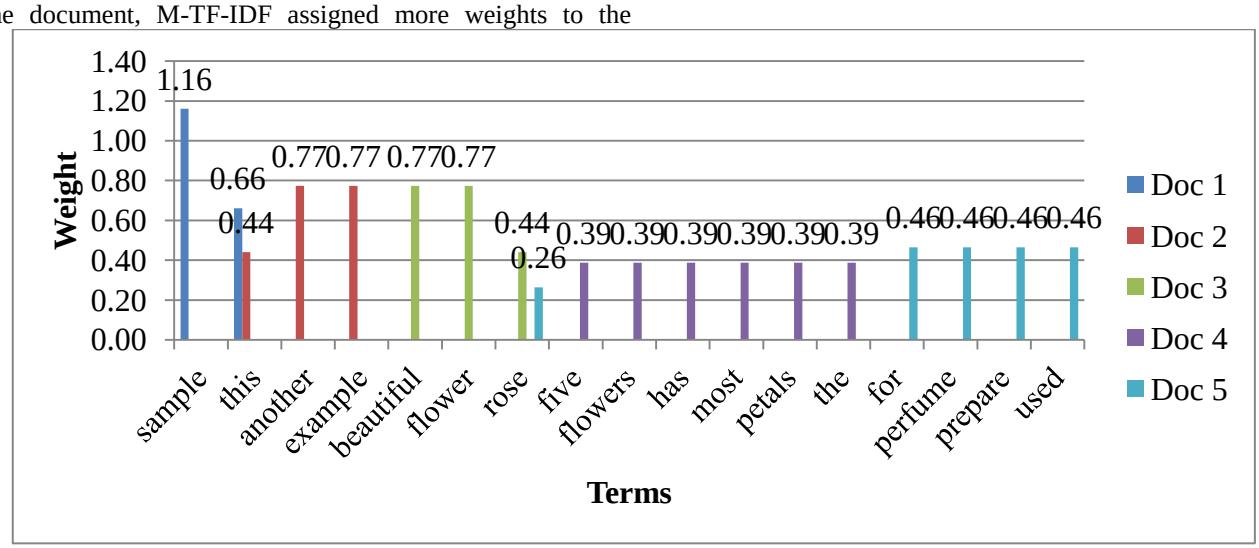

Fig. 1. Terms with weight assigned by TF-IDF method

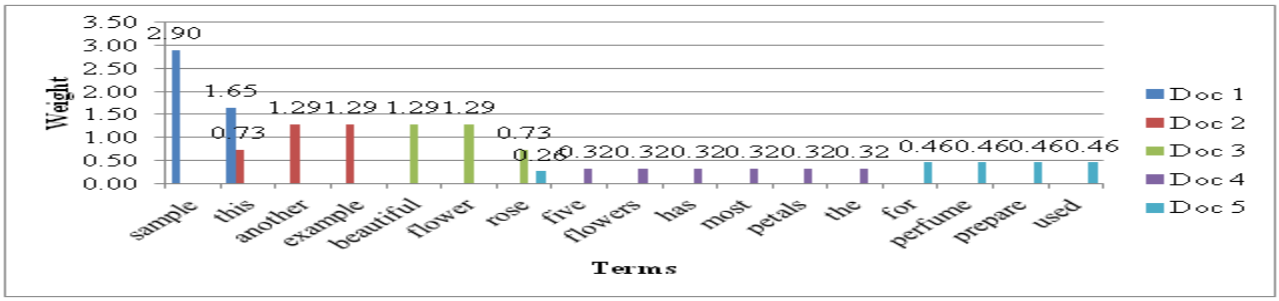

Fig. 2. Terms with weight assigned by M-TF-IDF method

In this work, the text document data set considered for processing is divided into training documents and test documents. The documents with class labels have been considered for preprocessing and feature extraction. During the feature extraction process, terms available in the document which is related to the labels are selected for further processing. Hence, the important words got more weight and the less important words got less weight than the previously assigned weight by the existing TF-IDF. This enlarged TF-IDF is used for extracting words and assigned appropriate weights to the terms which are important for further processing.

V. IMPORTANCE OF SUPPORT VECTOR MACHINE

Support Vector Machine is a supervised learning method. SVM is used for classification of text data for the past few decades for its best performing classification approaches. In today's machine learning applications, SVM is considered as the best one and which is most robust and gives accurate result when compared to other algorithms. In SVM, the kernel function and cost parameter(C) are important to produce high accuracy. In the training phase, the selection of kernel is also important. Kernel functions of SVM are linear, sigmoid, polynomial, Radial Basis Function(RBF) and exponential RBF. These kernel functions do not have the guarantee to produce good performance for all type of data sets. The data points are mapped onto the high dimension vector space using the kernel functions. Both the parameters are important to produce the most favorable hyper-plane. The non-linear classification differs from the linear classification in terms of kernel function. If the data is linear, linear kernel is used. If the data is non-linear, kernel functions sigmoid, polynomial, RBF and exponential-RBF are used. Each of the kernel functions has unique characteristics and performs well for different types of data. So, selection of kernel function is important to improve the classification accuracy. This process increases the processing time and computational cost.

Figure 3 shows the linearly separating optimal hyper-plane. In this figure, (○) represents one type of data, (△) represents another type of data. There are two hyper-planes separating the data points into two types in which, the optimal separating hyper-plane (denoted by the dark line on Fig. 3) select one which is closest to the data points of each type. The closest data point's dist1 and dist2 of each type are called as support vectors. The sum dist1+dist2 are the margin of the hyper-plane. Optimal hyper plane mainly depends on the kernel functions and cost parameter.

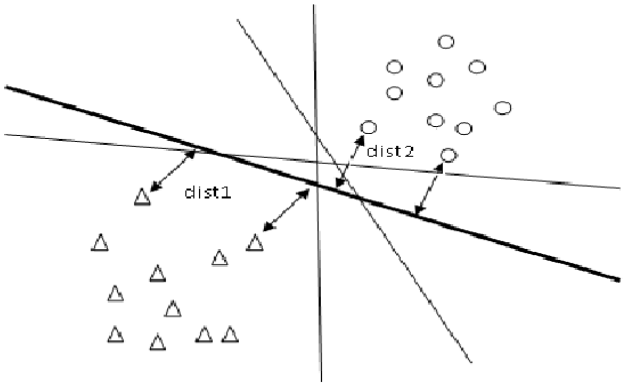

Fig. 3 Linearly separating optimal hyper-plane

VI. PROPOSED ALGORITHM FOR TEXT DOCUMENT CLASSIFICATION ALGORITHM

A. Frame work

The frame work of how to extract features from the text document is shown in the figure 4.

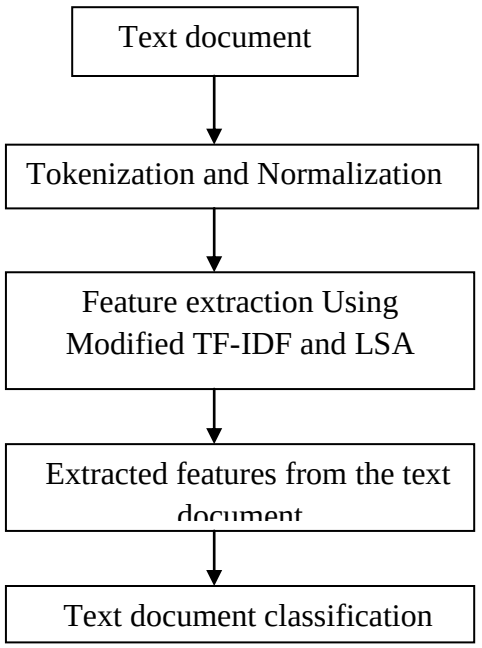

Fig. 4. Frame work of feature extraction procedure and classification

B. Algorithm

Classification algorithm:
Objective : Efficient and effective classification
Input : Collection of documents
Output: Documents are effectively classified
Pre-processing:

1.Reduce the size of the text document using tokenization and normalization

2. Construct Document-Term matrix

3.The text document transformation has been done in the following 4 steps,

i)Weight1=(Tt/NTD)*log(ND/NDt)* (ND/NTD)

ii) w = LSA(DTM)

iii) Weight2= as. textmatrix(w)

iv) X = Weight1+Weight2

Training:

1. This vector space X is given as input to the SVM.
2. SVM produced support vectors that are considered for the further process.

Testing :

1. To assign the label to the new test data point,

i) The average similarity has been calculated using the dot product versus each group support vectors and the test data.

ii) If the new test data has the highest similarity with any one group of support vectors, then assign the label of that group to the test data.

2. Generate the classification result.

VII. EXPERIMENTAL RESULTS

The proposed text classification has been implemented in R-software. The performance of the algorithm is evaluated with the four text corpus, namely Reuters-21578 R8 dataset, 20 news group dataset, descriptive type answer dataset and the crime news dataset.

Data set

Reuters-21578 R8 dataset

This is the benchmark dataset often used for text classification. This dataset is retrieved from the Reuters-21578 text categorization collection web site[12]. It consists of 7674 documents. These documents are categorized into eight groups. The training set consists of 5485 documents and the testing set consists of 2189 documents. The table II catalogs the Reuters-21578 R8 dataset labels.

Table II:Reuters-21578 R8 dataset labels

S.No.	Type
1	Acq
2	Crude
3	Earn
4	Grain
5	Interest
6	Money-Fx
7	Ship
8	Trade

20 news group data set

This dataset is one of the bench mark dataset for text classification. It consists of 18821 documents. It is divided into 11293 as training documents and 7528 as testing documents. This dataset is downloaded from the website [12]. Each document in the dataset consists of class label and sequence of words separated by spaces. This dataset has 20 different categories. From this group, for text document analysis only five categories such as alt.atheism, comp.graphics, sci.med, talk.politics.misc and rec.autos are selected.

Descriptive type answer dataset

This dataset is formed through collecting answers from the learners. Learners are grouped into under-privileged, normal and bright. This dataset consists of students answer. It consists of 420 documents. This dataset consists of six kinds of labels is shown in table III. It is divided into training and testing documents.

Table III. Descriptive type answer dataset labels

S.No.	Type
1	Five marks
2	Four marks
3	Three marks
4	Two marks
5	One mark
6	Zero mark

Crime news dataset

The news articles for this study collected from English news papers in and around Madurai region[13]. Collected articles are stored as separate documents. Documents are classified into two types: crime document and non-crime. The documents are split into two groups: Training documents and Test documents. The dataset consists of 790 documents of which 600 are used for training and 190 are used for testing. It is shown in the table IV.

Table IV: Crime news dataset labels

S.No.	Type
1	Crime
2	Non-crime

A. Comparison of existing TF-IDF and proposed M-TF-IDF feature extraction method

All the four datasets are preprocessed using tokenization and normalization methods. First Reuter-21578-R8 data set is considered for comparison. Tokenization method split the each text document in this dataset to tokens. The tokens are normalized by changed the letters of each word into small letters, removed numbers, removed the punctuation symbols, removed the stop words and produced stem of each word in the text document. Then apply the feature extraction method TF-IDF and M-TF-IDF to extract the important features from the document. Each methods discussed above are used to extract the features. Each feature has assigned weights.

Each dataset consist set of training and testing documents. The document term matrix(DTM) is constructed for each preprocessed document. The DTM is given as input to the individual feature extraction methods TF-IDF and M-TF-IDF. Each methods assigned weights to the extracted features from the text document. The weights assigned to the terms are considered for comparison. This process is also continued for other three data sets like 20 news group data set, descriptive type answer data set and crime news data set. After all the feature extraction process was completed for all the four dataset, the selected features weights are used for further processing. Features weight assigned by TF-IDF and M-TF-IDF are shown in the Table V and figure 5 for all the four datasets.

This comparison shows that the features weight assigned by using M-TF-IDF is better than the weight assigned by using TF-IDF. The proposed M-TF-IDF assigned more weights to the term which are more important and less weight to the term which are less are important. But this method does not consider the semantic similarity between the terms in the document. Hence, decide to use LSA method to extract features from the document.

Table V: Features weight assigned by TF-IDF and M-TF-IDF

Methods	Reuters-21578-R8 dataset	Crime news dataset	Descriptive type answer dataset	20 News Group dataset
TF-IDF	1054.345	2029.588	132.5681	9852.7
M-TF-IDF	3489.718	8631.509	310.2872	36434.5

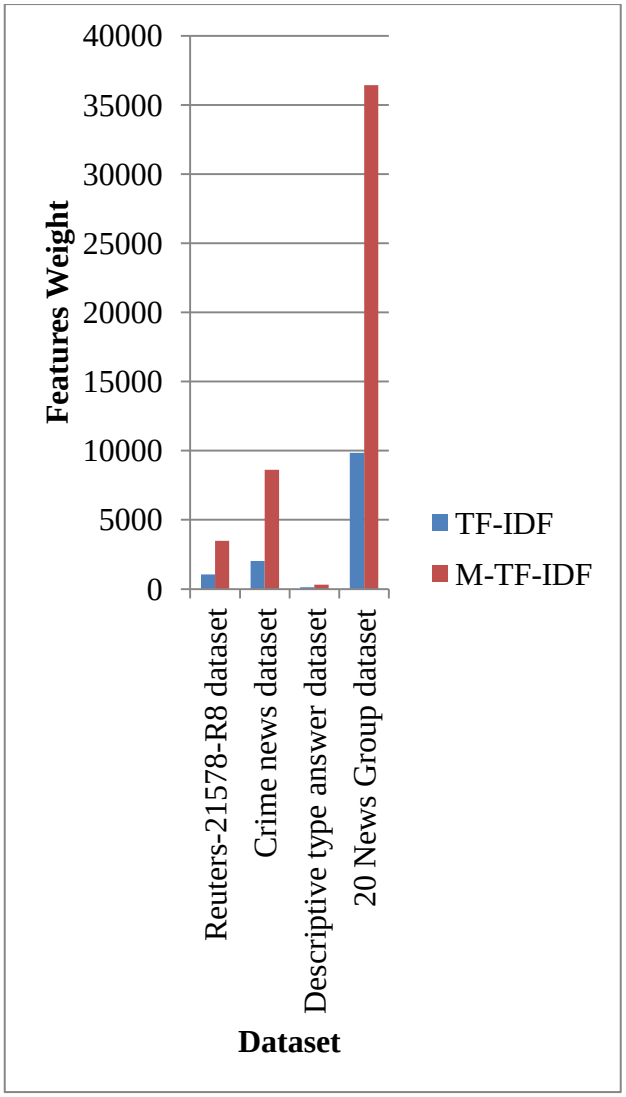

Fig. 5. Comparison of existing (TF-IDF) and proposed (M-TF-IDF) feature extraction method for datasets

B. Comparison of existing and proposed with semantic similarity feature extraction method

The proposed M-TF-IDF method only assigned weights based on the important of the term. To extract the features from the document based on the semantic similarity, Latent Semantic Analysis is used. First apply the pruning and stemming process for four data sets. The document term matrix (DTM) is constructed for each preprocessed document. The DTM is given as input to the feature extraction methods. Then apply the combination of feature extraction method TF-IDF with LSA and M-TF-IDF with LSA. The combination of proposed feature extraction methods M-TF-IDF and semantic similarity method LSA is compared with the combination of existing feature extraction TF-IDF and LSA for all four data sets. It is shown in the table VI and figure 6. This comparison shows that the performance of M-TF-IDF with LSA is better than TF-IDF with LSA methods in terms of the extracted features weight. This proposed method considers both the importance of terms and semantic relationship exists between the terms in the document. Extracted features by using proposed method used as input for classification using SVM.

Table VI: Features weight assigned by TF-IDF+LSA and M-TF-IDF+LSA

Methods	Reuters-21578-R8 dataset	Crime news dataset	Descriptive type answer dataset	20 News Group dataset
TF-IDF + LSA	25138.22	37347.02	7094.65	166915.4
M-TF-IDF+LSA	27573.59	43948.95	7272.369	193497.2

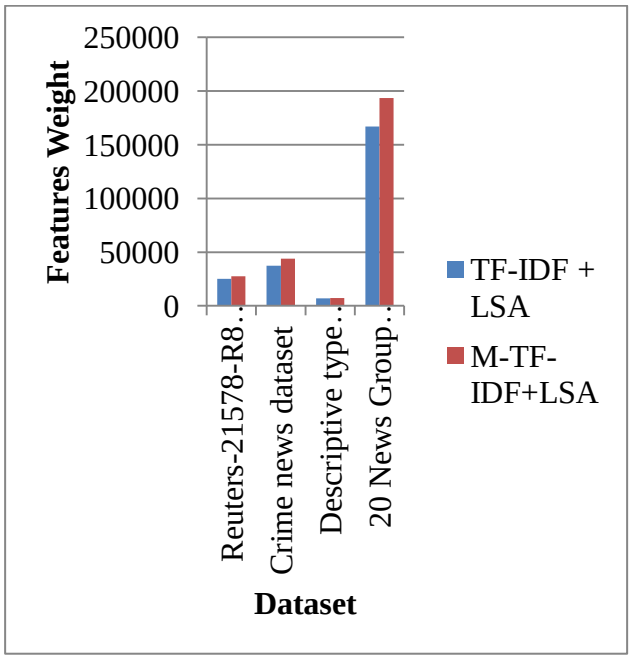

Fig. 6. Comparison of TF-IDF+LSA and M-TF-IDF+LSA feature extraction method for datasets

C. Analysis of performance of proposed algorithm

The proposed feature extraction method has been used to extract better features from the document. The extracted features have been given as input to the classifier SVM.

Support Vector Machine is used to classify the dataset. Four different types of datasets have been used for conducting experiments. The bench mark dataset Reuter-21578-R8 and 20news group dataset are used for analyzing the performance of the feature extraction methods. After examining the performance of the proposed method, it has been used to evaluate the descriptive type answer into 1 mark, 2 marks, 3 marks and 4 marks and also used to filter the news document into crime or non-crime. The classification performance of the proposed is shown in the table VII. The Reuter-21578-R8 dataset achieves 88.3% of accuracy and 20 news group dataset achieves 87.75% of accuracy. The real time dataset descriptive type answer dataset achieves 80.16% of accuracy and crime news document dataset achieves 72.14% of accuracy. The performance of proposed feature selection method is displayed in the figure 7.

Even though the existing feature extraction methods TF-IDF assigned weight to the terms, the proposed method M-TF-IDF performs well in the assignment of weights to the features extracted from the text document. The M-TF-IDF method assigns more weights to the important feature and less weight to the other features of document. When compares TF-IDF method with M-TF-IDF method, the performance of the M-TF-IDF has been best in terms of weight assignment. But it does not consider the semantic similarity between the terms. Hence, feature extraction method LSA is extract the semantic features from the text document and assign weights to the features based on the semantic similarity. The features space produced by the LSA is compressed using the mathematical technique Singular Value Decomposition (SVD). The extracted features from the LSA are also considered for comparison with existing and proposed feature extraction methods. The table VII and figure 7 and figure 8 showed the performance of proposed algorithm. This table shows that the proposed algorithm gives more accuracy than the other combination methods.

Table VII: Comparison of results obtained using proposed algorithm for four data sets

Dataset	Method	Accuracy
Reuter	TF-IDF+SVM	75.3
	M-TF-IDF+SVM	80
	LSA+SVM	77.67
	Proposed algorithm	**88.3**
20news group	TF-IDF+SVM	73.34
	M-TF-IDF+SVM	75.8
	LSA+SVM	74.12
	Proposed algorithm	**86.75**
Crime news	TF-IDF+SVM	64
	M-TF-IDF+SVM	67.61
	LSA+SVM	66.87
	Proposed algorithm	**72.14**
Descriptive type answer	TF-IDF+SVM	71.49
	M-TF-IDF+SVM	73.26
	LSA+SVM	78.23
	Proposed algorithm	**80.16**

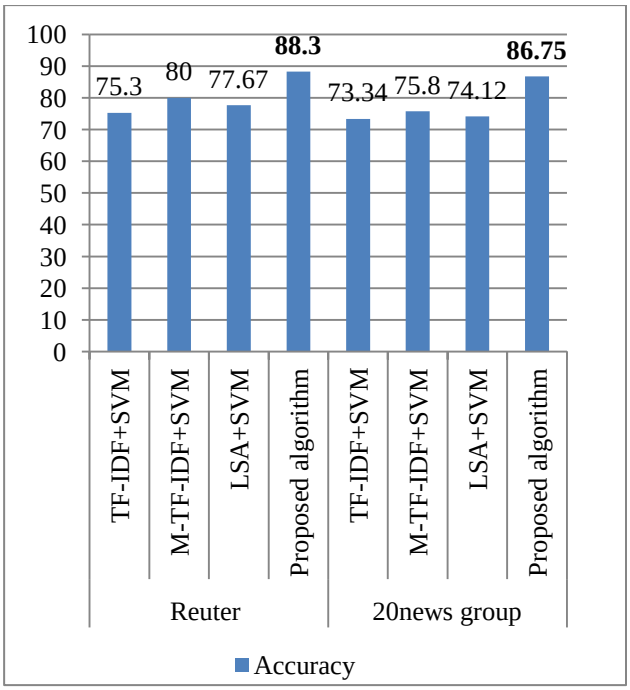

Fig. 7. Performance of proposed feature extraction methods of Reuter and 20 news group data sets

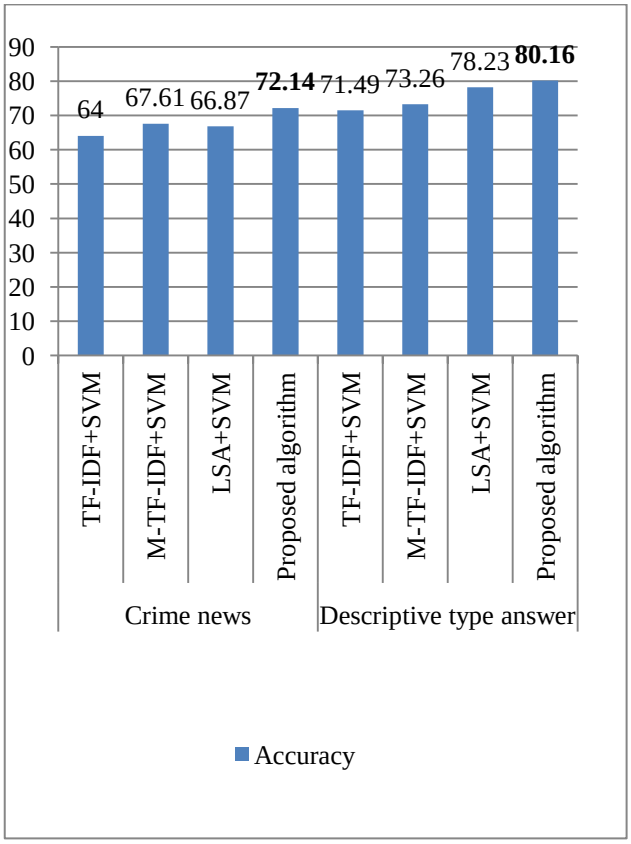

Fig. 8. Performance of proposed feature extraction methods of crime news and descriptive type answer data sets

D. Evaluation of descriptive type answer using proposed algorithm

Students' descriptive type answers are collected from the students of a class. Students are categorized into three classes, viz., are brilliant, average and under-privileged. The students' answers are collected and stored in a separate file. The collected answers are evaluated with the help of other staff, the result of which may contain deviation in some student answers. The proposed work tried to eliminate this deviation [14].

The collected answers are preprocessed using Pruning and Stemming processes. As a result of preprocessing processes, the articles, punctuation, affixes and suffixes attached to the words are removed. So the size of the file is reduced. Sample Students answers are displayed in Table VIII without any modification in the answer. Each answer is considered as one document. Document is saved in the name of the student roll number. After pruning and stemming, M-TF-IDF and LSA methods are used for selecting the informative words from the students answer. The words extracted from the students answers are displayed in the table IX. The documents after assigned the weight using M-TF-IDF+LSA are classified using SVM.

The descriptive type answers documents are represented using weights assigned by M-TF-IDF and LSA. These documents are classified using SVM. The documents nearest to the optimal separating hyperplane are selected as support vectors.

To assign mark to the new answer document, dot product has been calculated between the new answer document and the support vectors. Based on the result of dot product, mark will be assigned to the new document. Thus, the descriptive type answers are evaluated as 1 mark, 2 marks, 3 marks and 4 marks. Table X shows the selected students answer weight and their grade.

Students' answers are evaluated using this proposed algorithm. This algorithm is useful for academic institutions and universities for evaluating the descriptive type answers written by learners. This method has a number of benefits like increased reliability of results, reduced time and effort, reduced burden on the faculty and efficient use of resources. In future, online learning and teaching would be the essential one for educational institutions. This proposed algorithm uses semantic information present in the form of relations between words in sentences. It allows more accurate assessment and more effective evaluation of the learning process.

Table VIII: Student answers

S.No.	Answer
1	Comparing the Algorithms: Based on various Factors, user may choose the algorithms. Target ->Algorithms can generate association rules and these rules should satisfy support and confidence level of dataset. Type ->Algorithms can generate regular (or) advanced association rules. Datatype ->User can use common techniques such as apriori, sampling and partitioning algorithms to generate association rules. Itemset techniques -> Itemsets are to be counted in various ways. Transaction techniques ->Itemsets are to be counted by scanning the database of transaction. Itemset Datastructure ->All the techniques can use hash tree data structure to store the candidate itemset and the count. Transaction Data structure -> All the transactions are represented in a TID list. Architecture ->Algorithms are proposed in sequential, parallel and distributed architecture. Parallelism technique -> Algorithms uses data parallelism and task parallelism.
2	Comparision algorithm: Target: Algorithm should generate the association rules and these rules satisfy support and confidence level. Type: Algorithm may generate the regular or advanced association rules. Data type: Algorithm may generate the association rules for various types of data. Technique: User use the common technique such as sample, partitioning and apriori. Itemset technique:",Itemset can be counted in various ways. Transaction technique: For counting the itemset database of transaction scanned. Itemset Datastructure: All the algorithms use has 3 datastructure to store the candidate itemset and that count. Transaction technique: It represent in the TID list. Architecture:Algorithms are proposed by sequential and parallel and distributed datastructure. Parallelism: It represented in 2 ways Data parallelism. Task parallelism.

3	Comparing algorithm:based on the various factor user may choose the algorithmTarget:Algorithm may generate the all association rules &rules should satisfy the support and confidence Type: Algorithm may generate the regular (or) advanced rules. Data Type:Algorithm should generate the association rules for various types of data Technique: user use the common technique such as apiriori, sampling for generating the Association rules Itemset Technic: Item set can be counted in various ways Transaction Technique: Counting the itemset database of the transaction in scanned. Transaction datastructure: All the datastructure are represent in a TID list Architecture: Algorithm are proposed by sequential parallel and distributed architecture Parallelism technic: Algorithm use data parallelism(or)Task Parallelism"

Table IX: Terms selected from students answer

Compare	Algorithm	apriori
Confide	Datatechniqu	candid
User	Rule	generat
accocia	Use	Various
Type	Itemset	Transact
Parallel	Clount	Datasturcture
Support	Level	dataset
Regular	Base	Hash
Store	Task	Common
Tree	Target	Factor

Table X: Sample students answer grade

Roll number	Weight	Mark
14	95.12	4 marks
07	95.12	4 marks
52	91.29	4 marks
183	91.29	4 marks
244	82.55	3 marks
300	78.37	3 marks
274	74	3 marks
136	71.52	3 marks
254	62.69	2 marks
162	59.78	2 marks
263	56.22	2 marks
42	50.58	2 marks
55	43.89	1 mark
40	36.99	1 mark
261	35.38	1 mark

E. Classification of news as crime or not using proposed algorithm

News articles analyzing is one of the emerging research topic in the past few years. News paper discusses various types (political, education, employment, sports, agriculture, crime, medicine, business, etc) of news in different levels such as International, National, state and district level. In this news articles, crime discussion plays a major role because one crime leads to a many other crimes and also affect many other lives. In India, Madurai is one of the important places which have many historical monuments. Madurai is a sensitive place. The news articles for this study collected from English news papers in and around Madurai region[13].

This proposed algorithm analyzes the crime on Madurai region from the January 2015 to December 2015. Police departments face a number of crime problems because Madurai region is one of the sensitive areas in Tamilnadu. The distinctive crimes meet by the public, such as murder, theft, alcohol-related problems such as underage drinking and drunk driving, as well as assaults, sexual assaults, and rape. Such problems often consume many of their resources. This system can help to police departments in preventing such crimes in or around Madurai through the analysis of crime occur in previous year. The previous year analysis is used to determine why a problem was occurring, who was responsible, who was affected, where the problem was located, and what form the problem takes.

The primary objectives of the proposed algorithm on crime data set are 1) to select enlightening features related to crime 2) to train and test the SVM classifier model using the selected features with different kernel settings 3) to provide information to the police department about the analysis of crime in and around Madurai.

Crime data set is analyzed in different ways which are useful to the crime department. The words extracted from the news documents are displayed in the Table XI. The documents after assigned the weight using proposed feature extraction method are classified using SVM. Table XII shows the result of documents weight with the classification result.

The documents nearest to the optimal separating hyperplane are selected as support vectors. To classify the new document, dot product has been calculated between the new document and the support vectors. Based on the result of dot product, classified the document as crime or non-crime.

The following figure 9 shows that the number of crime occurred in the months of 2015 in and around Madurai. From that figure, crime department able to analyze the crime rate of months. This figure shows that June month has the highest crime rate and March month has the lowest crime rate.

The focus of this system reflects the fact that this type of analysis is used for police departments and that there is an increasing demand for the application of crime analysis that utilize data available from new papers. This analysis helps to police department to reduce the occurrence of crime in the future.

Table XI: Terms selected from crime news

Polic	Court	Case
Woman	Person	District
Hous	Kill	Gold
Madurai	Govern	Arrest
Road	State	Office
Found	Youth	Investig
Accus	High	Bench
Death	Vehicle	Depart
Member	Report	Order
Team	Work	regist

Table XII: Sample documents filtered as crime or non-crime

Document number	Weight	Crime/Non-crime
574	468.69	Crime
667	449.33	Crime
62	319.63	Crime
112	275.73	Crime
660	252.98	Crime
864	210.68	Crime
918	189.14	Crime
318	97.97	Non-crime
403	96.66	Non-crime
565	90.63	Non-crime

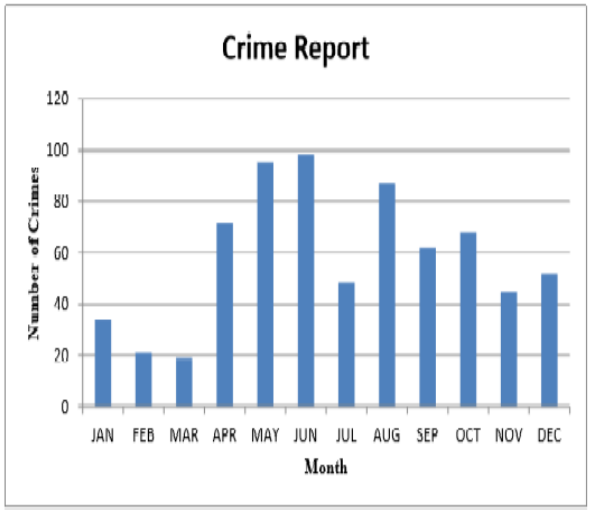

Fig. 9 Number of crimes in the year 2015 in and around Madurai city

VIII. CONCLUSION

The proposed text classification algorithm preprocesses the document. Then the proposed feature extraction method used to select the features from the text documents. The modified TF-IDF used to assigned weight to the features based on the occurrence and most importance. LSA used to extract features based on the semantic similarity value. The combination of M-TF-IDF and LSA with SVM provides high accuracy. The performance of the proposed and existing method is compared using four datasets. This comparison shows that the proposed feature extraction method performs well than the existing feature extraction method in terms of weight assignment to the extracted features of the text document. The performance of the proposed algorithm has been tested with bench mark data set. It shows better performance for the bench mark data set. Hence, this proposed algorithm is used for both evaluate the students descriptive type answers' and also for classify the crime data set. This proposed work used to evaluate the students descriptive type answer. The discrepancy in the manual evaluation is reduced by using this proposed method. This more effective evaluation process is used to increase the results reliability, efficient use of resources, reduced effort and time taken for evaluation of descriptive type answer. The news documents collected in and around Madurai are also filtered using this proposed method. Police departments face a number of crime problems because Madurai region is one of the sensitive areas in Tamilnadu. This system can help to police departments in preventing such crimes in or around Madurai through the analysis of crime occur in previous year. The proposed system used to select enlightening features related to crime then train the SVM classifier model using the extracted features. Then identify the new document whether it is crime or non-crime and provide useful information to the police department about the analysis of crime in and around Madurai.

This proposed algorithm uses SVM for classification. The SVM classification accuracy mainly depends on the kernel functions and the cost parameter. This dependency has been removed in future.

REFERENCES

1. Uysal, Alper Kursat, and Serkan Gunal. "The impact of preprocessing on text classification." Information Processing & Management 50.1 (2014): 104-112.
2. Bullinaria, John A., and Joseph P. Levy. "Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD." Behavior research methods 44.3 (2012): 890-907.
3. Moh'd A Mesleh, Abdelwadood. "Chi square Feature extraction based SVMs Arabic language text categorization system." Journal of Computer Science 3.6 (2007): 430-435.
4. Zareapoor, Masoumeh, and K. R. Seeja. "Feature extraction or feature selection for text classification: A case study on phishing email detection." International Journal of Information Engineering and Electronic Business 7.2 (2015): 60.

5. Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "A comparative study of TF* IDF, LSI and multi-words for text classification." Expert Systems with Applications 38.3 (2011): 2758-2765.
6. Jivani, Anjali Ganesh. "A comparative study of stemming algorithms." Int. J. Comp. Tech. Appl 2.6 (2011): 1930-1938.
7. Willett, Peter. "The Porter stemming algorithm: then and now." Program 40.3 (2006): 219-223.
8. Jing, Li-Ping, Hou-Kuan Huang, and Hong-Bo Shi. "Improved feature selection approach TFIDF in text mining." Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on. Vol. 2. IEEE, 2002.
9. Meena, K., and R. Lawrance. "Semantic similarity based assessment of descriptive type answers." 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). IEEE, 2016.
10. Danielle S.McNamara(2010). "Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition", Topics in Cognitive Science.
11. Sidorov, Grigori. "Latent Semantic Analysis (LSA): Reduction of Dimensions." Syntactic n-grams in Computational Linguistics. Springer, Cham, 2019. 17-19.
12. http://www.cs.umb.edu/~smimarog/textmining/datasets/
13. K.Meena and R. Lawrance. "News document analysis by using a proficient algorithm", Int. Journal of Engineering Research and Application, ISSN : 2248-9622, Vol. 7, Issue 6, (Part -2) June 2017, pp.07-13.
14. K.Meena and R.Lawrance. "Text classification algorithm (TCLS) using Support Vector Machine", Journal of Advanced Research in Dynamical and Control Systems, ISSN: 1943-023X, Vol. 9, Sp-18/2017, pp. 1068-1090.

AUTHORS PROFILE

Dr. R.Lawrance has received B.Sc. & M.Sc. degree in Computer Science from St.Joseph's College, Trichy in 1993 &1998, M.Phil. Computer Science from M.S. University in 2003 and a Ph.D. degree from the Vinayaka Missons University in 2011. He has joined Ayya Nadar Janaki Ammal College since 1998 as an assistant Professor in B.C.A department. Now he has been promoted as Director in the Department of Computer Applications in the same college. His current research interest lies in data mining and machine learning Algorithms. He has produced 18 M.Phil. Scholars and guiding for 8 Ph.D. Scholars. He has published 20 National level conferences, 30 International level conferences and 16 International level Journals.

K.Meena received her Bachelor degree in Computer Science from Sri Parasakthi College, Courtallam in 1998, Master of computer applications from Sri Saratha College, Tirunelveli in 2001 and Master of philosophy from Periyar University, Salem in 2008. She is currently doing research in Computer Science at Bharathiar University, Coimbatore. She published paper in 2 international level conferences and 3 in international level journal