Online News Text Classification using Neural Network and SVM

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Raghvan Gachli
(Roll No. 801232018)

Under the supervision of:
Dr. V.P. Singh
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2014

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*Online News Text Classification using Neural Network and SVM*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Mr. V.P. Singh* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

(Raghvan Gachli)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Mr. V.P. Singh)
Assistant Professor
Computer Science and Engineering Department

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

First of all, I would like to thank the Almighty, who has always guided me to work on the right path of the life. Due to mercy of God, it has been possible for me to reach so far. This work would not have been possible without the encouragement and valuable guidance of my supervisor Mr. V.P. Singh, Assistant Professor, Thapar University, Patiala. I thank my supervisor for his time, patience, discussions and valuable comments. I am equally grateful to Dr. Deepak Garg, Associate Professor and Head, Computer Science and Engineering Department, for motivation and inspiration that triggered me for the thesis work. I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable.

I will be failing in my duty if I don't express my gratitude to Dr. S. K. Mohapatra, Senior Professor and Dean of Academic Affairs the University, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

Raghvan Gachli

(801232018)

Abstract

Categories for classification of text are predefined according to these categories all text data is classified. We require classifying text to manage and search any data in database. There are many techniques available in market to classifying the text. Now days every website has overloaded text in database as like customer support websites, news website etc. so in this type of websites text need to classify. In news websites it's necessary to maintain record of old and new news into the database. The news can be classifying on the basis of predefined categories of type crime news, sports news, election news, entertainments etc. every technique that exists in real like SVM, Naive Bayes, and Neural classifiers, working well at a level with some limitations. In this we are going to discuss about these techniques and conclude with the comparison of results find out which technique can perform well. Text category detection refers to identifying the type of category getting used by the text. The process involves two process training and testing. The training section involves the feature extraction process and the testing section involves the identification of the type of text used. In the process classification, involvement of a classifier is there to check the accuracy of the training. In this we focuses on the enhancement of the text category detection using back propagation neural network and Support vector Machine. The classification results have improved by 5 to 10 percent

Table of Content

List of Figures

Chapter 1

Introduction

Text categorization or text classification is the topic in information science. It is a field which deals with the analysis, storage, collection, classification, categorization, retrieval, and manipulation etc. In today's era of globalization company are in the need of automatic classifying and categorizing the text documents. Companies are growing at a faster rate so as there databases. To classify the data present in the databases they need a automatic classification system. Automatic text classification begins in the early 1960 but with the immense availability of the online documents and the internet in the last two decade it regains its interest in the researchers. Before they used heuristic approaches/methods i.e. based on some expert knowledge the task are solved by applying some rules. But this approach is insufficient and now the focus is on fully automatic learning, classification and clustering methods. Text classification is basically assigning text documents and dividing them into different categories. It consists of two phases:

- Training phase
- Predicting phase or Testing phase

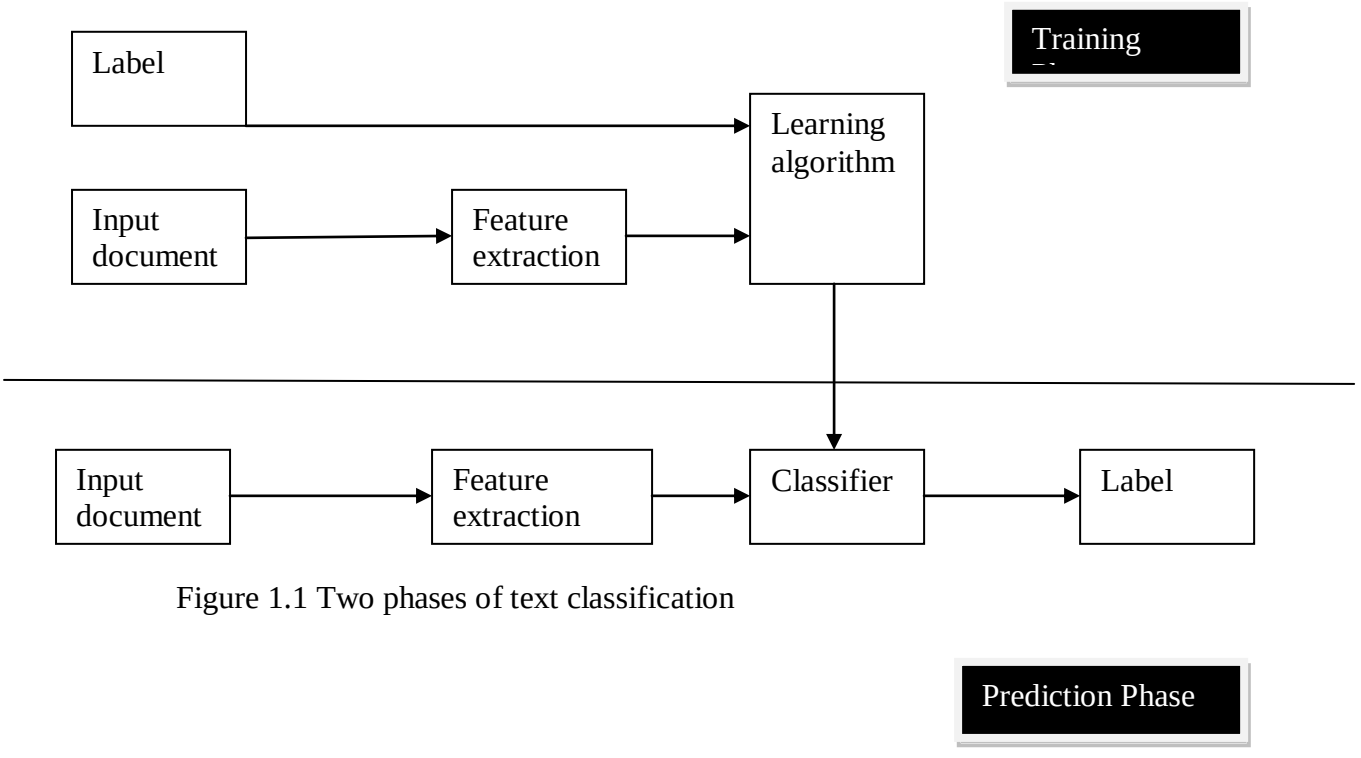

Figure 1.1 Two phases of text classification

Documents can be classified with respect to their subjects. There are two main philosophies of the subject classification of the document. First is the content based classification and the second is request based classification. Classifying the online news and analyzing the state of a person by his\her written document is challenging as well as essential for today's time. It is important because people online most of the time uses different, intellectual, slang etc words so as to describe, what they want to say about a particular subject. The other major purpose behind is to make the news more interesting. The general idea for classifying the text documents is that sometimes the expressions are not direct and sometime the meaning of concepts is different. All the websites are full of text databases like customer care support, FAQ, news etc. and it is necessary to classify then into their respective categories so that one can judge it properly. Specially in news have to maintain a large amount of data and also a need to classify the news into old and new news of different categories like crime, sports, economy etc.

1.1. Content based and Request based classification

In content based classification the weights are given to some particular subjects in the document which will determine the class of the document in which that should be assigned. Whereas request based classification the anticipated request from the users decide the class of the document and influence the document classification.

1.2. Automatic Document Classification

It is divided into three parts:

- Supervised learning Document Classification
- Unsupervised learning Document Classification
- Semi-supervised learning Document Classification

1.2.1. Supervised learning

It is a machine learning task in which from a labeled training data a function is surmise. In supervised learning the training data is a group of examples and each example is a pair of an input object and an output value. The training data is analyzed through its learning algorithm and a function is produced, which

can be used for mapping. The unseen instances are labeled into different classes thorough an optimal scenario.

1.2.2. Unsupervised learning

It is the task to find out the hidden layer in the unstructured /unlabelled data. As the data is unlabelled, so there is no error to evaluate the solution. It is also related to density estimation problem like in statistics. It has many techniques that tells about the key features of the data and summarize the data. Many methods of this learning are based data mining mostly used to pre-process the data.

1.2.3. Semi-supervised learning

It is the combination of both supervised and unsupervised learning. It uses the task and techniques of supervised learning and also uses unlabelled data of the unsupervised learning. It produces better results when some unlabelled data is used in conjunction with a small amount of labeled data.

From a theoretical point of view, both the supervised and unsupervised learning only differ in their casual structure. In supervised learning one set of observation (inputs) has a cause on another set of observations (output). In unsupervised learning all the observations assumed that they are the cause by latent variables.

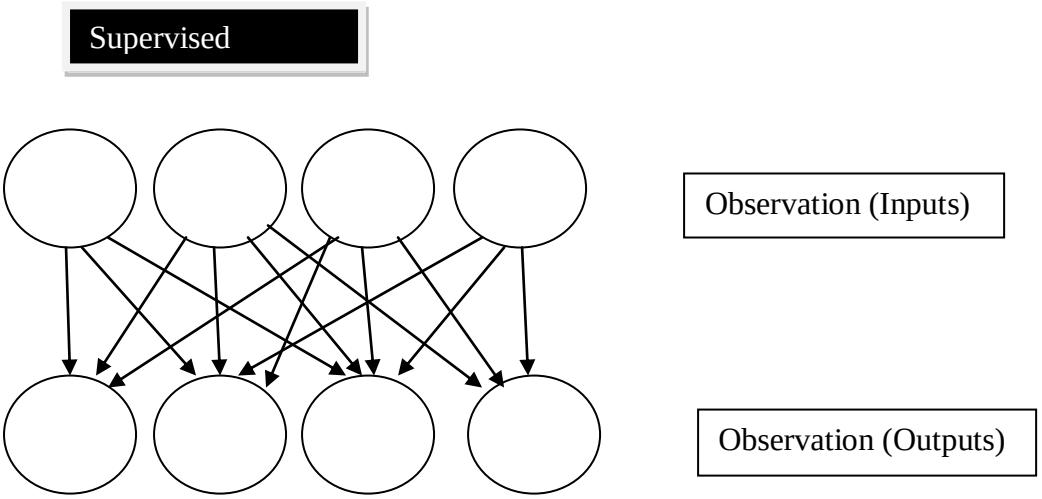

Figure 1.2 Structure model of supervised learning [20].

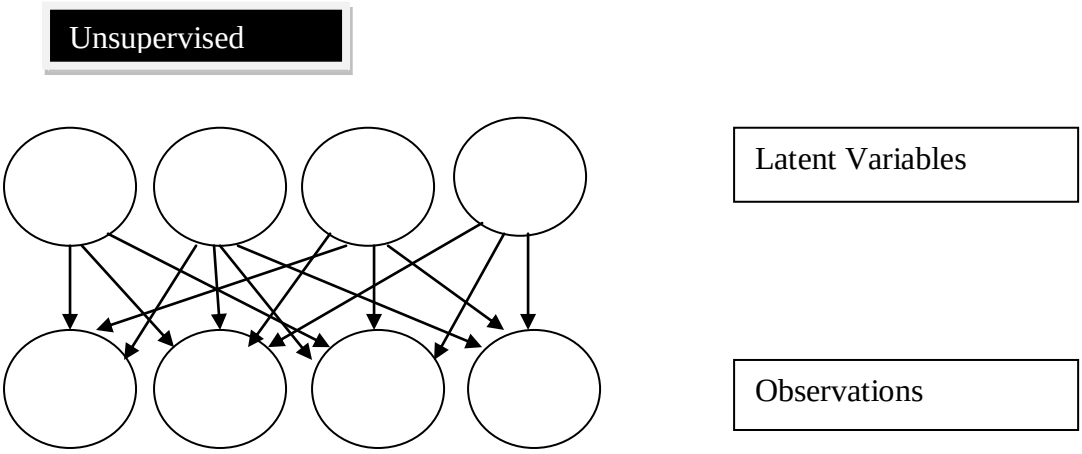

Figure 1.3 Structure model of unsupervised learning [20].

1.3. Data Mining

Data Mining is a process for searching patterns in large databases. The Mining of the data means searching out a piece of data from a bulk data block. As we are fetch knowledge and hence this is also called as the knowledge discovery in databases. (KDD) [1]. The basic work in data mining is divided into two parts. First is classification and the other is clustering. Even though all refer to different kind of same area but still there is difference in both the material. Clustering is the task in which the set of objects grouped in such a way that similar are more similar becomes a cluster and dissimilar objects form another cluster. It may serve as the pre processing step for other algorithms, such as characterization, attribute subset selection, and classification, which operates on the detected clusters and the selected attribute [2]. If the clusters are identified then the classification of the data is possible. By applying the searching algorithm one can find out the maximum number of clusters in a specified region.

Classification is based on two following parameters:

- An area which is used for the classification that is the cluster region.
- Type of dataset to be applied on the selected region.

1.4. Terms Related With Data Mining

1.4.1. Data

Data is a form of facts, numbers, or texts that is processed by a computer system. Now days, large amount of data in different formats and different database such as.

- Transactional data - sales, cost, inventory, payroll, and accounting etc..
- Non-operational data - industry sales, forecast data, macro economic data etc.
- Data about the data itself i.e. Metadata, such as logical database design or data dictionary definitions

1.4.2. Information

The patterns, associations, or relationships belong to given data provides information. For example, analysis of retail point of sale transaction data provides information on which products are selling and when.

1.4.3. Knowledge

Information converts into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior.

1.4.4. Various types of Data Mining

- Association Rules
- Classification
- Clustering
- Sequential Patterns
- Sequence Similarity

1.4.4.1. Association Rules

They are used to find out the common /casual relationships in the data base. They are used in Store layout, catalog design, customer segmentation etc.

1.4.4.2. Clustering

There are some given n points; to separate them into k clusters is called clustering. It is used for information retrieval; identify similar web documents, mapping the universe etc.

1.4.4.3. Sequential Patterns

They are used to find out the frequently occurring patterns in a given set of events. It is used for analyzing a given set of data, medical diagnosis etc.

1.4.4.4. Sequence Similarity

In this we look for the similar trends in a given number of data sets. It is used in finding stocks, geological irregularities etc.

1.4.4.5. Classification

A set of rules to partition the data into different /separate group. It is used for classify people, weather prediction, fraud detection, variation etc.

Possible solutions:

- Bayesian classification
- Neural Networks
- Genetic Algorithm
- Decision Trees
- Support Vector Machines
- Expectation Maximization (EM)
- Term frequency –inverse document frequency (tf-idf)
- Latent Semantic indexing (LSI)
- Natural language processing (NLP)

1.5. Navies Bayes or Bayesian Classification

It is a group of simple probabilistic Classifier by applying the bayes theorem between the features with strong independence. It is a popular method of text classification. It work on judging the frequency of the words and partition it into one or more categories. Because only the independent variables are taken into consideration so it has an advantage small amount of data is required (training) to estimate the parameters that are taken into account.

1.6. Neural Networks

Neural Networks is one of the most advanced classifiers in the testing category. The neural has a feed forward method. The feed forward method takes one input as a training sample and another input as the target sample. The input sample is the data stored in the database on the basis of all the features which have been extracted at the time of training.

If P is the input sample then P is defined as

P=sum (all features (input));

In the same manner, there would be the testing feature or the target sample. The target sample are the scenario which would be fixed and are provided as an input. The general architecture has been presented as below.

The figure 1.4 illustrates the general working principle of the Neural Networks.

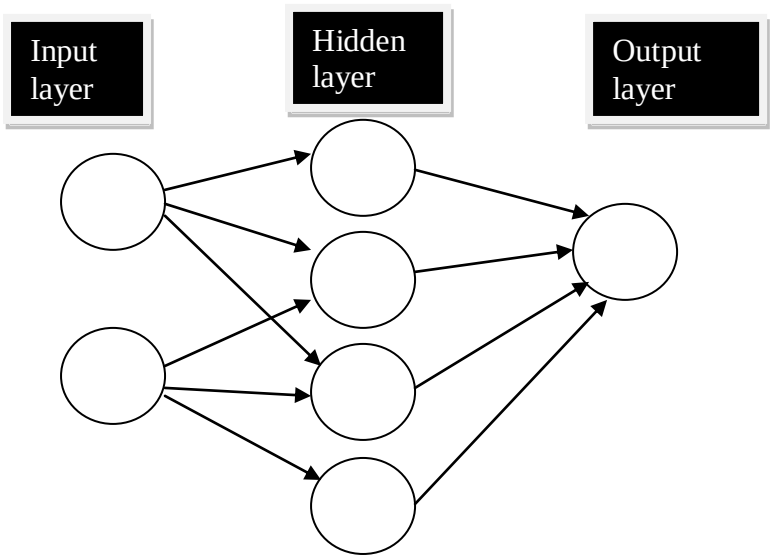

Figure 1.4 Neural Network Model.

Here the central block is termed as the Neural Classifier. There are two input samples where the neural classifier generates the weight accordingly for the first input which has been taken from the database. The second input is the target set which is to be tested. The neural takes each sample as a neuron and explains to the architecture that how the input is going to react and how the result is going to be proceeded. Finally it produces a binary result. If we do precede more than one sample category, the neural will have be combined.

1.7. Genetic Algorithm

In the field of artificial intelligence is a search for finding a solution for the problem when the classic methods are unable to find the exact solution. It is a process of natural selection. It is a class of evolutionary algorithms to find solutions for the problems using techniques like inheritance, mutation, selection, and cross-over which are inspired by the process of survival of the fittest or in general the natural selection.

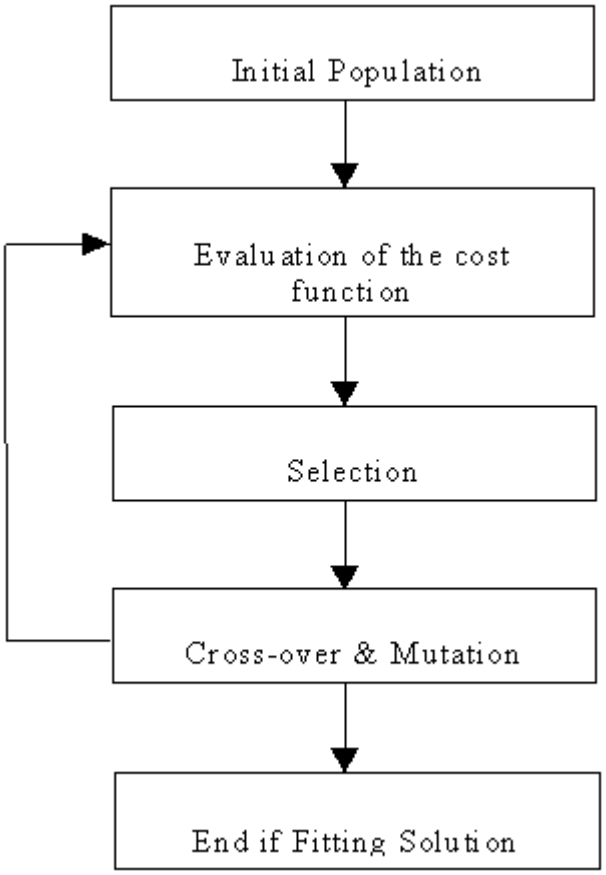

Figure 1.5 Flow chart of Genetic Algorithm [20].

1.8. Decision Tress

It is used in decision analysis. It is a decision support tool that gives tree like graph. It includes resource cost, outcomes, utilities etc. It is a flow chart like structure in which there is one internal node on which test is performed on some attributes. Every outcome of this test became a branch like structure and reaches the leaf node which represents the class label i.e. the final decision taken by taking all the attributes into computation. The way from root node to the leaf represents the classification rule. Three types of nodes are there in decision nodes, chance nodes, and end nodes. Decision nodes are represented by squares. Chance nodes represented by circles. End

nodes represented by triangles as shown in figure 1.6. They are commonly used in operation search to reach a goal result.

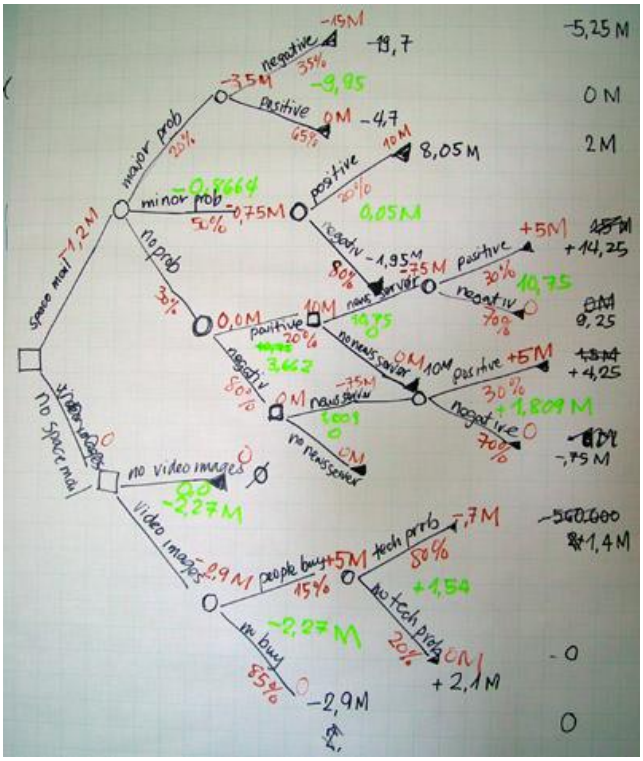

Figure 1.6 Decision tree [19].

1.9. Support Vector Machine

Support vector machine classifier is used to make segments of selected data on the basis of emotions and simple text. Input data is presented in two sets of vectors in n-dimensional space, a separate hyper-plane is constructed in space due to which margin between two data sets maximize. There are many hyper-plane exist for classify data but we have to find that hyper-plane which provide maximize margin between two data sets. Due to set of support vectors, risk for structure minimizes. This is n-dimensional hyper-plane where define n is number of features of input vectors, that is necessary to define boundary of classes. Binary classes are required to classify training data. Main goal of performance for SVM is to minimize risk of structure. By a training example as given (a1, b1), (a2, b2) …. (a_n, b_n), positive and negative examples separated by hyper-plane. If a point a1 exists on hyper-plane then it satisfies: (w. a1) + d = 0 where d represents distance from origin and w is normal to hyper-plane. There is shortest distance between negative and positive examples define by margin of hyper-plane. A kernel function is used to fit data on hyper-plane. We

cannot directly fit data on hyper-plane without SVM mechanism. User provides a function like a line, polynomial which select support vector along surface of this function. “Curse of dimensionality” is main property that is used to avoid upper bound on VC-dimension. VC-dimension is used to measure capacity of the machine [29]. As shown in fig 1.7.

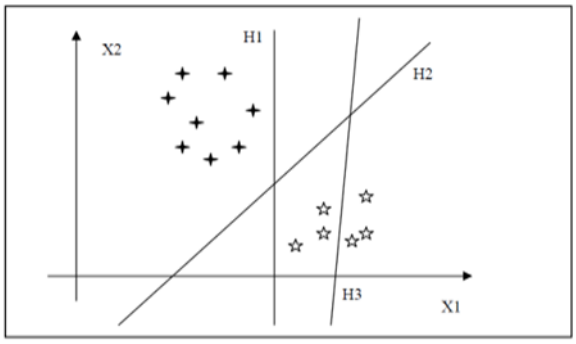

Figure 1.7 An example of SVM [20].

Main object of SVM is try to find nearest distance between point of same class and maximize with point of other class and draw hyper-plane in two categories very clearly as possible.

1.10. Expectation-Maximization (EM)

The Expectation-Maximization algorithm is a part of statistics. An iterative method used to find the maximum likelihood parameters in the stat model in the case where the equations cannot be solved directly. These models include latent variables, unknown parameters in addition with some known observations i.e. either the data has some missing points or the model can formulate it by assuming the existence of the unknown data. To find out it requires derivatives of the likelihood function with the unknown data. The parameters and the latent variables both solve the equation at the same time. If talk about the statistical models it is not possible with the case which has latent variables but the result comes out to be a group of interlocking equations. In the solution the latent variables values are requires by the parameters and vice versa but by interchanging the equations into one another produces an equation that cannot be solved.

1.11. Term frequency-inverse document frequency (td-idf)

It is also a part in statistics. It tells about that how much important a word in a document or in a group of documents. It is mostly used as a weighting factor in text mining. If a word appear in the document most of the time then tf-idf value helps in increasing the proportionally. Moreover it also helps us to generalize that some words are more common than the other. Its weighting factor is used mostly by the search engines for scoring and ranking the documents present online. It is mostly used for stopping words filtering.

1.12. Latent Semantic Indexing (LSI)

It is a method of indexing and retrieval. It uses a mathematical method named singular value decomposition (SVD). This method is used to identify the patterns, relationships between terms and concepts in the unknown data. It works on the principle the words used for same context tend to have similar meaning. The main ability of the LSI is to find out the contextual concept from the text through some associations with the term that occur in similar context. A word that has many meaning and many words that has the same meaning, this type of problem is solved using LSI [25][22].

1.13. K-Nearest Neighbors

K-Nearest Neighbors are used for classification and regression. It is a non parametric method. The input consists of k-nearest training examples. It has a feature space in which the input resides and the output is depend upon for which k-NN is used either classification or regression.

1.14. Natural Language Processing

It is concerned with the interactions between human and computers through a natural language. General task include in NLP [22] is automatic summarization, co reference resolution, discourse analysis, machine translation, morphological segmentation, named entity recognition and many more. The main objective of NLP is that we should make the computer understand through human or some natural language. There are many challenges in NLP and one of the major challenges is natural language understanding as it is the interaction between human and the computers.

Chapter 2

Literature Review

Some Researchers presented the news in Web Page Classification Method (WPCM). Neural network model has been used with inputs obtained by the principal components and Class Profile-Based Features (CPBF) [1]. Regular words with fix numbers from each class will be used as a feature vectors. These feature vectors are used as the input to the neural networks for classification. WPCM provides acceptable classification accuracy for the datasets of sports news. The other researcher also proposed the automatic text classification method to assign text files to one or more predefined categories according to the text information files. SVM is recognized as one of the most effective text classification methods for its high accuracy [2]. A new rule extraction method for text classification based on trained SVMs is proposed to solve the bottleneck of SVMs. The approach they used can improve the validity of the extracted rules either in speed or accuracy. Used kernels in SVM looks at continuous data, and neglects the structure of the text. Classical kernels, has been proposed for use of various string kernels for spam filtering. Data pre-processing is a vital part of text classification, feature vectors are generated by SVM kernels [3]. The feature mapping variants in text classification (TC) are used to improved performance for the standard SVM in filtering task and an online active framework for spam filtering has been developed. The web classification mining system has been described using support vector machine. The ability of pattern recognition, self-learning and generalization of SVM. New classification mining method is explained to classify the web text information. The results of this study shows the feasible and effective for Web mining classification is used to classify the web information. Genetic algorithm and support vector machine are used here [4]. Support Vector Machine (SVM) in data mining is a classification technique, based on structural risk minimization principle and VC theory using statistical theory [5]. An intelligent system for online news classification based on Hidden Markov Model (HMM) [6] and Support Vector Machine (SVM) [6] they focuses on improving the speed of the system when the data to be computed is huge. An intelligent system is designed to extract the keywords

from the online news paper and classify it according to the pre defined categories. Three different stages to classify the online newspapers based on (1) Text pre-processing (2) HMM based Feature Extraction and (3) Classification using SVM. Data have been collected for experimentation from different newspapers. HMM used for feature Extraction. SVM used for Text Classification [6]. Through introducing the basic principle of SVM, they described further proposed a SVM-based classification model. SVM is an effective machine learning method [7]. The text classification problem is high dimensionality of the feature space. Its study shows that the combination of GA and k-means algorithm is quite useful in reducing the high feature dimension, improved accuracy and efficiency for text classification. Genetic algorithm and k-means algorithm can select relevant features in text classification [8]. A method for text classification; it is an important part of text mining. The task is then to determine a classification model that can assign the correct class to a new document in domain. It involves Document pre-processing, Feature extraction / selection, Model selection, Training and testing the classifier and depicts that SVM is outstanding from other with its effectiveness to improve text classification [9]. A pre-defined category group with the proper training set based on the activation of FPI and attempted to classify the document using FPI methodology. The algorithm involves are text tokenization, text categorization and text analysis algorithm are explained for the classification of the documents [10]. The different rule-based classification approaches in data mining are described for the Arabic text categorization. “If-Then” knowledge in order to decide the most applicable one to Arabic text classification problem has been explained. The rule-based classification algorithms such as: One Rule, rule induction (RIPPER), decision trees (C4.5), and hybrid (PART) has been discussed. The results indicate that the hybrid approach of PART achieved better performance when compared other algorithms [11]. Many documents are available in digital forms which need classification of the text to solve this problem machine learning techniques has been used. The main benefit is less use of expert work and straightforward portability to different domains are possible. Other researchers examined the main approaches to text categorization comparing the machine learning paradigm [12]. The text Categorization is a pattern classification task for text mining and necessary for efficient management of textual information systems. The documents can be classified by three ways unsupervised, supervised and semi supervised methods. This presents a comparative study on different types of

approaches to text categorization and different types of approaches: .K. Nearest Neighbor, Decision Trees, Naïve Bayes Algorithm [13]. The feature vector of the document are grouped into clusters, proposed using Fast Fuzzy Feature clustering. The numbers of iterations required to obtain cluster centers are reduced. Principle Component Analysis with slit change is used for dimension reduction. This method improve the performance by significantly reducing the number of iterations required to obtain the cluster center and was verified with three benchmark datasets [14]. In this an important algorithms that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved. Performance is evaluated on the basis of [15]:

- Precision [wrt ci (Pri)]is defined as the as the probability that if a random document [dx] is classified under ci, this decision is correct. Analogously, Recall [wrt ci (Rei)] is defined as the conditional that, if a random document [dx] ought to be classified under [ci], this decision is taken
- TPi–The number of document correctly assigned to this category.
- FN - The number of document incorrectly assigned to this category
- FPi - The number of document incorrectly rejected assigned to this category
- TNi - The number of document correctly rejected assigned to this category
- Fallout = FNi / FNi + TNi
- Error = FNi +FPi / TPi + FNi +FPi +TNi
- Accuracy = TPi + TNi

As the use of internet is increasing day by day and with the advancement of internet news also publish online. So to handle this bulk amount of news various data mining techniques for classification had been used. In this paper we are using an intelligent system based on Hybrid algorithm (HMM, SVM and CART) [16] for e-news classification. An intelligent system is designed which will extract the online news and then will find out category and subcategory wise news for 2 categories. In future it can be extended to other sub-categories [16].

- Used HMM for text classification.

- Used SVM for text classification.
- Used Hybrid approach using HMM, SVM and CART for text classification.

Many researchers have combined the different classifiers in order to get better document of text Classification results. In this they combined k-NN and SVM [17] although they produced good results but has limitations like if the training data size increases SVM works degraded and if the attributes increases k-NN requires more processor, physical memory and takes more time to compute the results. Social networking sites like face book and twitter are majorly analyzed for online news these days. SVM and Navies bayes are combined [18]. Many classifiers are combined to get better and better results. There are other more things are required by the system in order to classify the text document into different categories. For example like keyword pattern matching, stop words removal, multiword etc. the keyword pattern matching The pattern matching problem is described as the problem of finding occurrences of keywords from in a given string [23]. In the context of category detection this method is based on certain predefined keywords. These words are classified into different categories such as disgusted, sad, happy, angry, fearful and surprised etc. process of keyword spotting method is shown in figure 2.1.

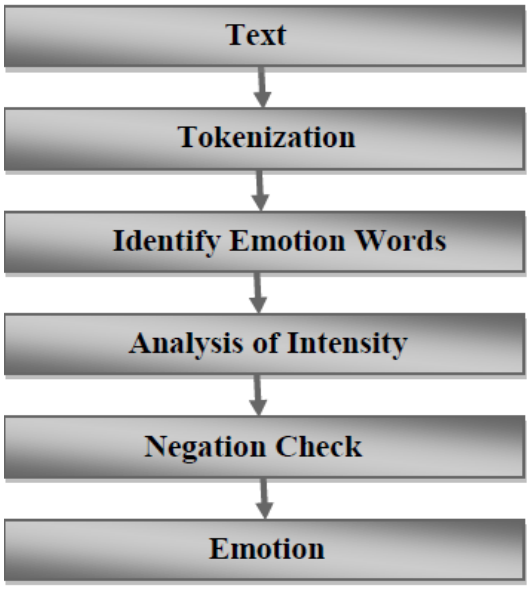

Figure 2.1 keyword spotting technique [23].

Keyword spotting technique for category recognition consists of five steps. A text document is considered as input and output is generated as a category class. The data is converted into tokens, from these tokens words are identified as category words.

Some text as input should be given and perform tokenization to the input text. Words related to categories will be identified. Afterwards analysis of the intensity of category words will be performed. Sentence is checked whether negation is involved in it or not then finally a category class will be found as the required output. Generally we have large vocabulary and text documents uses it but it is not necessary that all the words in the document are useful so the researchers proposed TF-IDF [24] LSI [25] and multiword [26] feature reduction techniques like and there combinations. Exhaustively and specificity [24] are two major things when documents classification is taken into consideration. Exhaustively describes the index and specify describes one index term. TD-IDF is a statistical method used to describe the importance of a word by evaluating how many times that word has occurred in the document and LSI and multiword and semantic methods covers the problem of polysemy and synonyms. Meta data plays a important role in text classification extracting the name of person, places, titles, authors name etc are some important keywords [27]. Getting information from heterogeneous HTML resources. Using a hierarchical neural network is better than the single neural network and categorical neural network [28] but only when the data is huge and consist of many categories or when many computers are used to train single neural network simultaneously.

Chapter 3

Gap Analysis and Problem Statement

It is all about classifying the online news section. The previous researchers have done a lot to achieve different targets. But there are loop holes in the previous research work. As per the previous research is concerned, it is all about classifying the basic categories of the news and inner categories using HMM, SVM and CART for 2 categories only. The previous research work does not mention anything about classifying the inner section of the major categories using HML, SVM and Genetic Algorithm and there combinations.

About 80% of the stored information is in text form. Text mining is still to having a high research potential. Knowledge can be gathered from many sources of information; still, unstructured texts remain the largest readily available source of Knowledge.

3.1. Gap Analysis

Based on the literature review of the concerned research papers the following gap analysis has been found.

- Researchers uses different algorithms to classify the text but there is no such consolidated approach for this particular statement. [15] [7] [18].

3.2. Problem Statement

We have used two algorithms for classifying the online news. First is the back-propagation neural network and second is support vector machine. Through the study of the previous research papers we found that for this particular case there is no such combination. They uses different classifiers like neural network, navies bayes, decision tress etc. individually as well as their combinations to classify the text documents. Through the regressive study of the literature review for this particular case this combination has not been used till now.

3.3. Problem Formulation

The previous researchers have done a lot for the text classification in online news classification but they have a least amount of work in terms of the internal structure. If we talk about the news category there is no Classification of e-POLITIC, e-FINANCIAL and e-SPORTS news. This combination for text classification has not been used till date to classify the online news. Researchers have done a lot on social networking (twitter, facebook) news and news from other sites like Yahoo and Google. It is time-consuming task to select the most interesting one as there was no proper classification of news articles. So to compute the classification results we will compare those values with real values to check the accuracy.

However, the problem deals with:-

"A Hybrid approach to classify categories of Online News using NEURAL and SVM"

3.4. Objective

The proposed objectives for the research work are: -

1) To study and analyze the existing Neural and SVM techniques for text classification.
2) To implement Neural and SVM based classification for categorization of online news.
3) To evaluate the performance of proposed model on the basis of various parameters: -
    A) Classification accuracy of Neural Network Model.
    B) Classification accuracy of Hybrid (Neural Network Model and SVM).

3.5. Tools and Platform

For many years, the main language for all engineering and scientific applications involving number crunching was C, C++ and Java. However, there are many things that cannot be done on these high-level languages. Cleve Moler developed MATLAB in 1970s. is a high-level language and interactive environment that enables you to

perform computationally intensive tasks faster than with traditional programming languages such as C, C++, and Java. In MATLAB the computations were carried out on whole matrices or vectors at once. Later in 1984, Jack little rewrote MATLAB in C incorporating more functionality including plotting gestures and founded The Math works Inc. to market it. Today MATLAB is a standard tool for both professional and academic use. In fact, for a million engineers and scientists in industry and academia, MATLAB is the language of technical computing. More than 400,000 technical professionals at the world's most innovative technology companies, government research labs, and financial institutions and at more than 2,000 universities rely it.

3.6. Functions Used

3.6.1. Newff

It is used to create a feed forward back propagation network.

Syntax

- net = newff(P,T,S)
- net = newff(P,T,S,TF,BTF,BLF,PF,IPF,OPF,DDF)

Description

- newff(P,T,S) takes,
- P - RxQ1 matrix of Q1 representative R-element input vectors.
- T - SNxQ2 matrix of Q2 representative SN-element target vectors.
- Si - Sizes of N-1 hidden layers, S1 to S(N-1), default = [].
- (Output layer size SN is determined from T.) and returns an N layer feed
- forward backprop network.

- newff(P,T,S,TF,BTF,BLF,PF,IPF,OPF,DDF) takes optional inputs,
- TFi - Transfer function of ith layer. Default is 'tansig' for
- hidden layers, and 'purelin' for output layer.
- BTF - Backprop network training function, default = 'trainlm'.
- BLF - Backprop weight/bias learning function, default = 'learngdm'.
- PF - Performance function, default = 'mse'.
- IPF - Row cell array of input processing functions.
- Default is {'fixunknowns','remconstantrows','mapminmax'}.

- OPF - Row cell array of output processing functions. Default is {'remconstantrows','mapminmax'}.
- DDF - Data division function, default = 'dividerand';

and returns an N layer feed-forward backprop network.

3.6.2. svmtrain

It trains a support vector machine classifier

Syntax

SVMSTRUCT = svmtrain (Training,Y) trains a support vector machine

(SVM) classifier on data taken from two groups. Training is a numeric matrix of predictor data. Rows of Training correspond to observations; columns correspond to features. Y is a column vector that contains the known class labels for Training. Y is a grouping variable, i.e., it can be a categorical, numeric, or logical vector; a cell vector of strings; or a character matrix with each row representing a class label (see help for grouping variable). Each element of Y specifies the group the corresponding row of Training belongs to. Training and Y must have the same number of rows. SVMSTRUCT contains information about the trained classifier, including the support vectors, that is used by svmclassify for classification. svmtrain treats NaNs, empty strings or 'undefined values' as missing values and ignores the corresponding rows in Training and Y.

3.7. Basic of MATLAB

MATLAB is a mathematical scripting language that looks very much like C++. Some features of the language are:

- Efficient matrix and vector computations.
- Easy creating of scientific and engineering graphics.
- Application development, including graphical user interface building
- Object-oriented programming.
- Extensibility (various Toolboxes)
- File I/O functions
- String Processing

Whenever, we create an M-file, we are actually writing a computer program using the MATLAB programming language. MATLAB commands are themselves M-files, which can be examined using type or edit command in MATLAB. The function files in MATLAB have a particular format, described below:

Function [output_args] =function_name(input_args)

There are 15 fundamental data types (or classes) in MATLAB. Each of these data types in the form of any array. This array is a minimum of 0-by-0 in size and can grow to an n-dimensional array of any size two dimensional versions of these arrays are called matrices. All of the fundamental data types are circled in the diagram below. Additional data types are user defined, object-oriented user classes (a subclass of structure) and java classes that you can wise with the MATLAB interface of java.

By default any numeric variable is assigned the data type double Matrices of type double and logical may be either full or sparse. For matrices having a small number of nonzero elements, a sparse matrix requires a fraction of the storage space required for an equivalent full matrix sparse matrices invoke special methods especially tailored to solve sparse problems. The logical data type represents a logical true or false value using the numbers I and 0 respectively. MATLAB returns logical values from its relational (e.g.,>, ~=) and logical (e.g., &&, xor) operations and functions. The char data type holds characters. A character string is nothing but a 1-by-n array of characters. You can use char to hold an m-by-n array of strings as long as each string in the array has the same length. Numeric data types include signed and unsigned integers, and single and double precision floating point numbers. However, all MATLAB computations are done in double prevision and to perform mathematical operations and integer or signal precision arrays, they must first be converted to bobble precision using the double functions. Structures and cell arrays provide a way to store dissimilar types of data in the same array. A Cell array provides yet another storage mechanism for dissimilar kinds of data. This is more intuitive as the data different data types here are stored in similar manner as in any matrix.MATLAB data types are implemented as classes. MATLAB gives provision to create user defined classes, which are nothing but subsets of the structure data type because these classes inherit from the structure class. A Java class is a MATLAB data type. There are built-in and third classes that are already available through the MATLAB interface. It is possible to also create Java class definitions and bring them into MATLAB Programming.

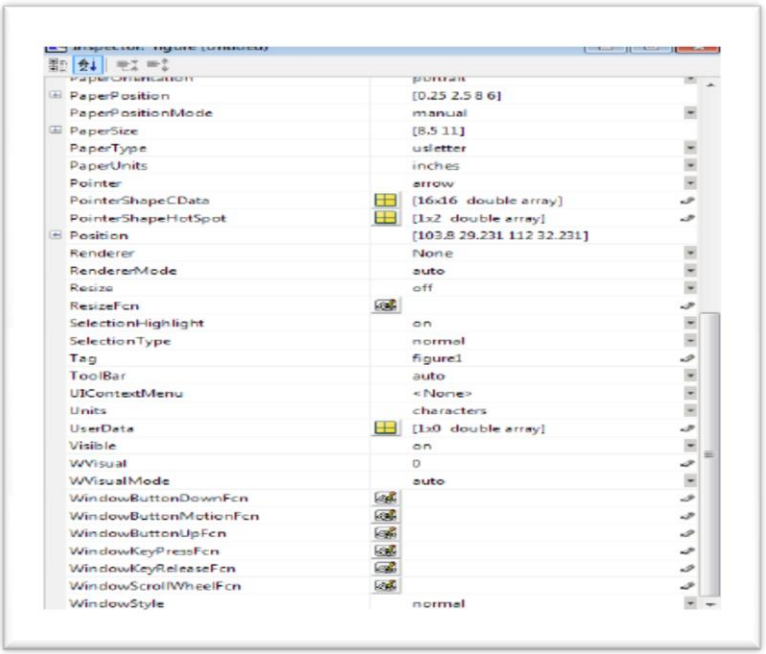

Figure3.1 Property Inspector Window

3.8. Algorithm

- Upload Files for all categories (sports , entertainment, politics, financial)
- Store values in db
- Target(1:4- for every category)
- Y=Net.Train (Uploaded_Vaues_files , targets, epochs)
- Upload a value for testing
- Test Sample=Input Sample
- G=Newff (Y, Uploaded Set,10) where 10 is the number of neurons
- If G<=1
- CATEGORY-sports
- Else if 1<G<2
- CATEGORY entertainment
- Else if2< G <3
- CATEGORY politics
- Else if2< G <3
- CATEGORY financial
- Call Neural-SVM for classification and plotting accuracy graphs

Chapter 4

Implementation and Experimental Results

The behaviors of the classifiers are very similar across the classification tasks. The performance depends on the number of documents in the train set. We require classifying text to manage and search any data in database. There are many techniques available in market to classifying the text

4.1. Snapshots

Figure 4.1 represents the main work window of the work done on the matlab. It contains two parts. In the first part on the Left hand side is the training part. In this we can manually add words of choice into the database. As we can see from the figure is that we have for sections on the left hand side section part i.e. Entertainment, Sports, Politics and Financial. By choosing the region of the choice adds a word into the text box. The right side section is the testing part with a small training part. By choosing a region of the choice on the right hand side we can upload a full document rather than putting single-single word. And we can test a document with neural only and the combination of both neural and SVM. And can have their accuracy graph.

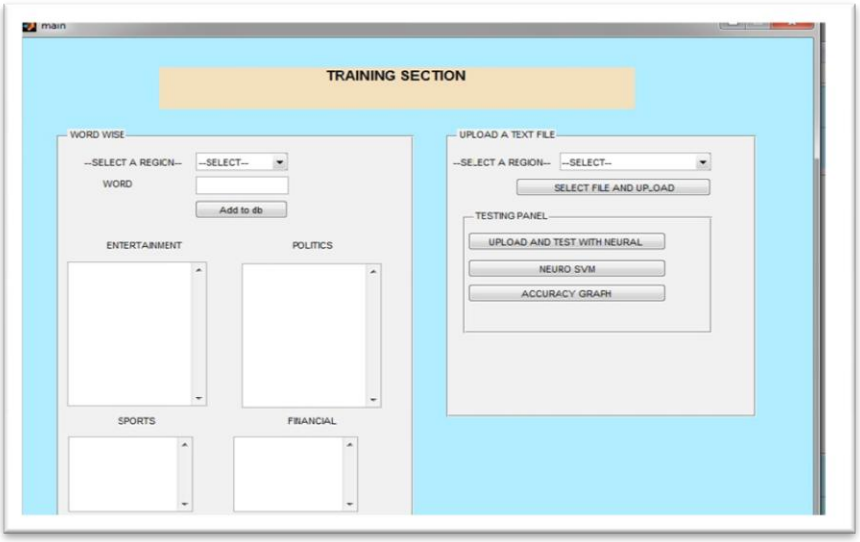

Figure 4.1 Main work window.

In figure 4.1, displaying the two sections training phase on the left side and testing phase on the right side the left side has 4 predefined categories (sports, politics, sports, finance). The right side has testing phase where the document can be tested and can be classified into the predefined categories.

Figure 4.2 shows the four regions or the four pre defined classes. Select the region of choice from the four classes and put the word into the text box.

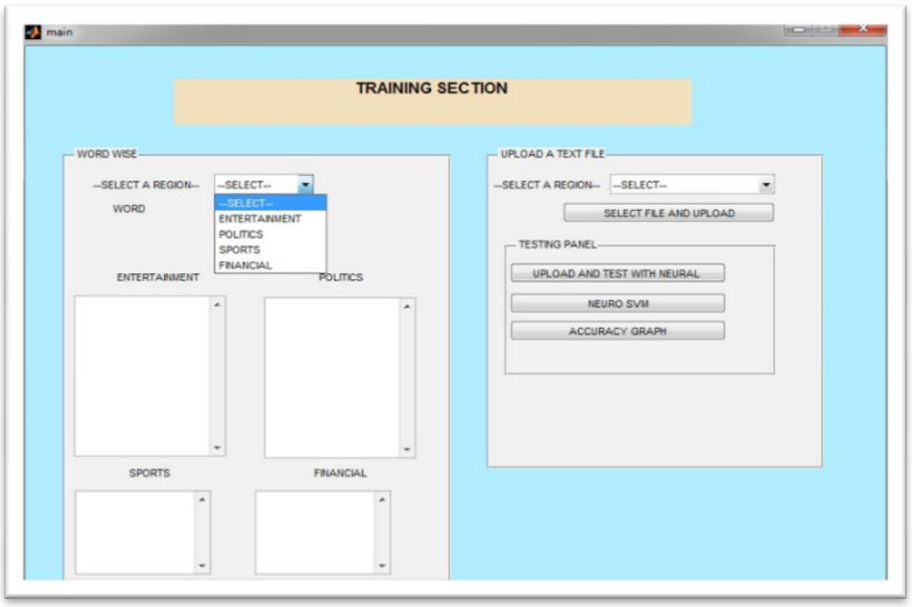

Figure 4.2 Pre-defined Classes.

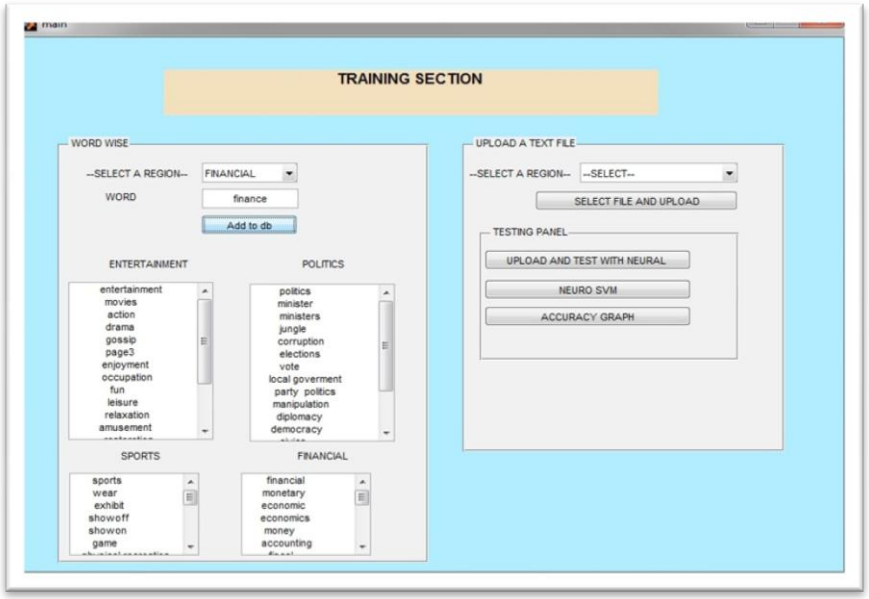

Figure 4.3 Words in the databases

In figure 4.3, shows the words put manually into the database of the four classes respectively.

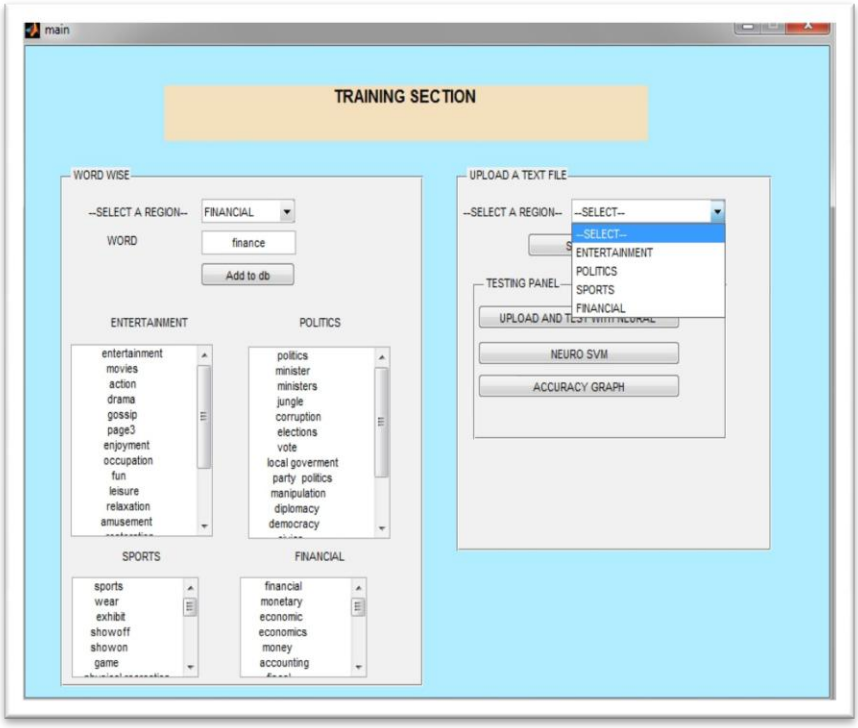

Figure 4.4 Upload a text file.

In figure 4.4, it is in the testing part of the system. Instead of putting word individually we can upload a file directly from here. And all the words present on that file will be updated into the database of the selected class.

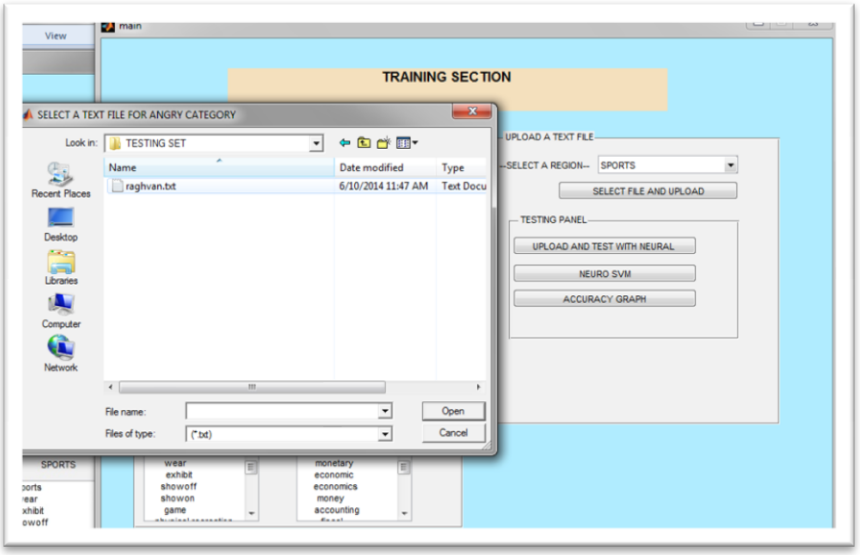

Figure 4.5 Testing model.

Figure 4.5, shows that after choosing the region of your choice. Click on the select file and upload button. As u click that button a small window pops up which has some files in .txt format. Double click on the .txt file and all the words will be updated into the chosen pre-defined class. As here it will update the data into the sports class.

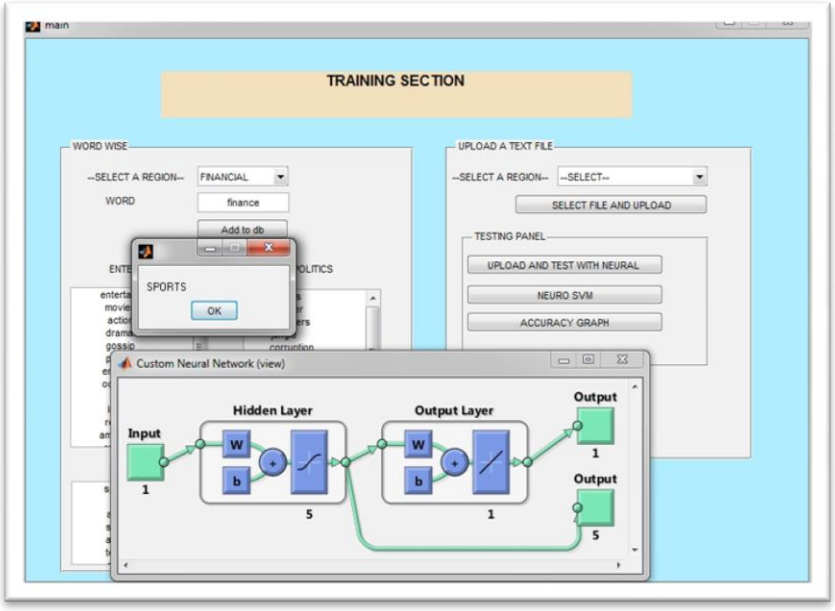

Figure 4.6 Simulation model for Neural Network

Figure 4.6 represents the results with only neural network. The figure also displays the custom neural pattern. In the figure we have uploaded all the databases and put a testing file to it and it distinguishes it into one of the predefined categories.

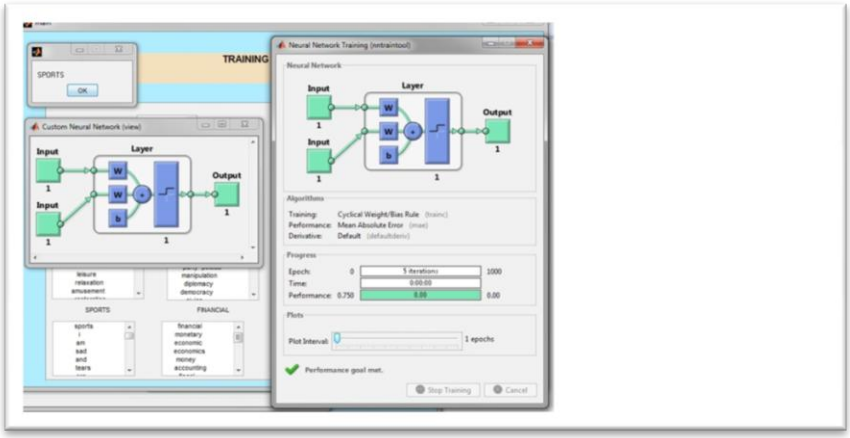

Figure 4.7 Simulation model for Neural-SVM.

In the figure 4.7 tells about the result from the hybrid approached used here. The combination of neural network and support vector machine. The figure here represents here the custom neural network diagram.

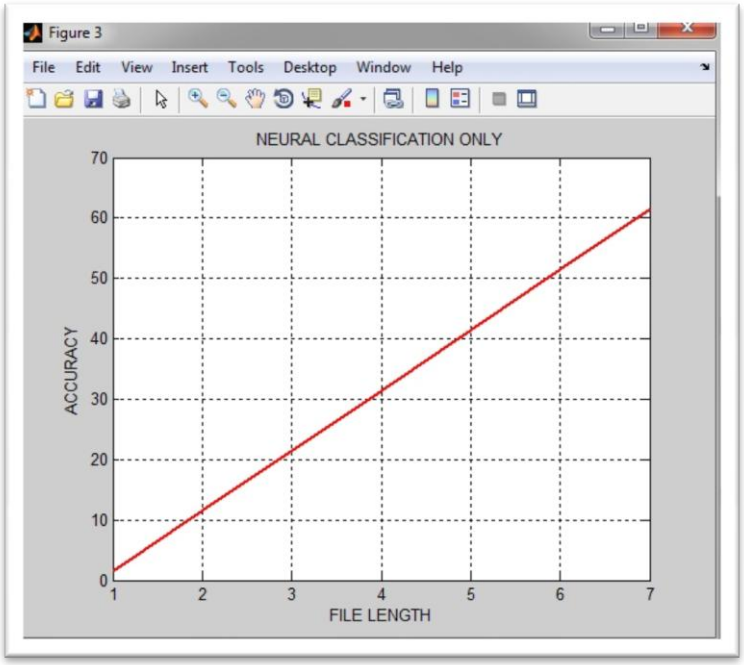

Figure 4.8 Classification accuracy graph of neural only.

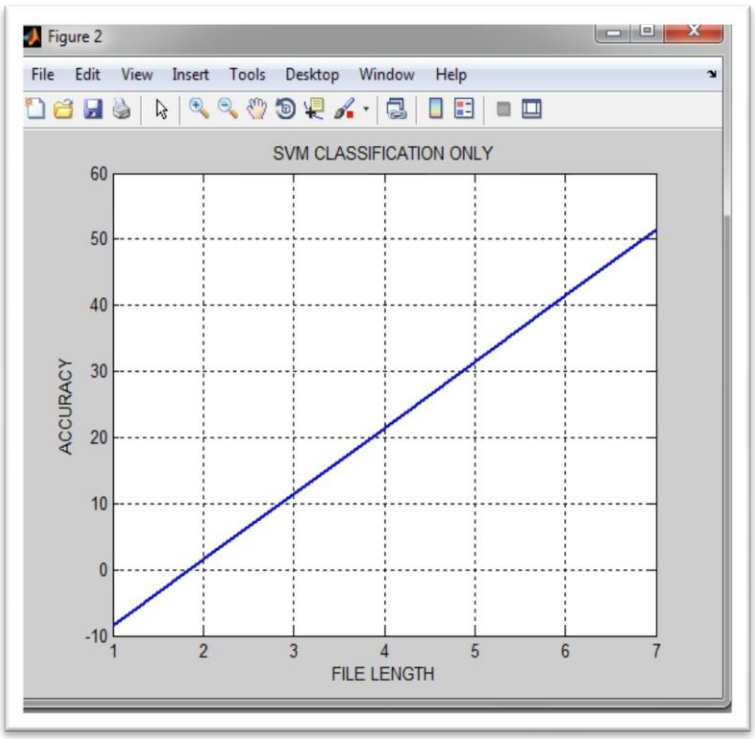

Figure 4.9 Classification accuracy graph of SVM only.

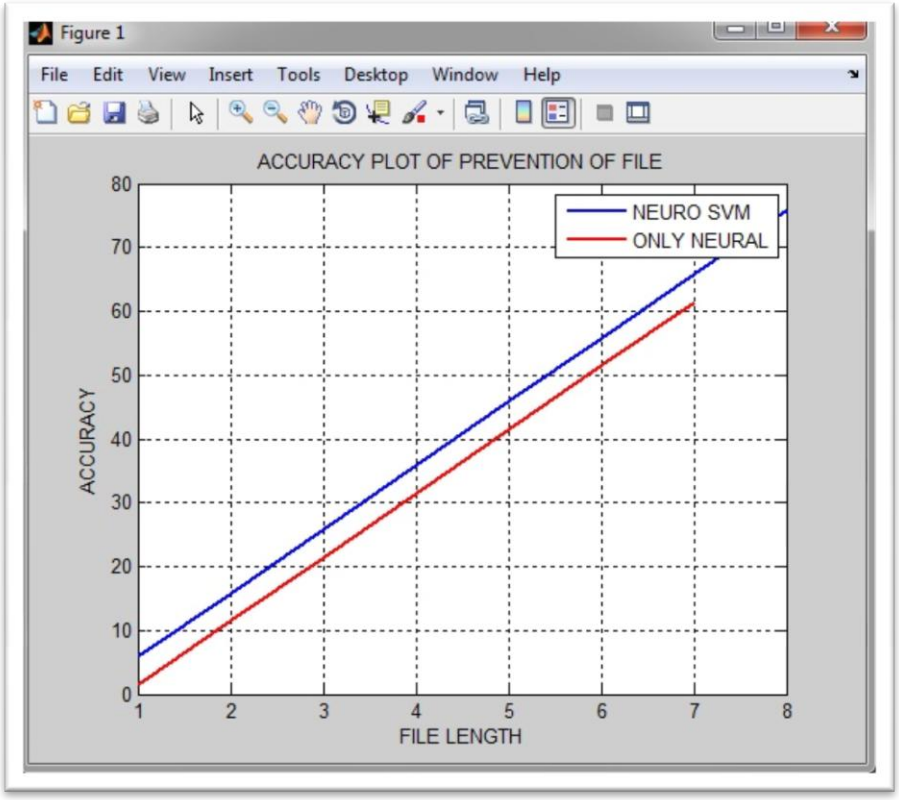

Figure 4.10 Classification accuracy graph of Neural-SVM (hybrid-approach).

Figure 4.8, figure 4.9 and figure 4.10 shows the classification accuracy graphs for neural network only, SVM only and the hybrid approach respectively used here. For the above figures it is clear that the accuracy of neural only is 62%. For SVM it is 52% and the hybrid approach is 76% when used together. The neural network weights are combined with input set provided. It works with hidden neurons. Another algorithm which has been put into action is Support Vector machine which binaries the entire system and takes the value as a binary input. The figure 4.9 represents the output of SVM only. It becomes quite clear that neural has a better performance edge over SVM. This research work as said has combined NEURAL with SVM. The accuracy of the hybrid classifier (Neural-SVM) which contains the neural network and the SVM classification. The results shows a good difference in the percentage growth of accuracy by almost 20 % when they are combined. We test another set of data randomly with both the neural and the Neural-SVM and it is classify into the politics class. The accuracy graph of the data is in figure 4.11, 4.12 and 4.13.

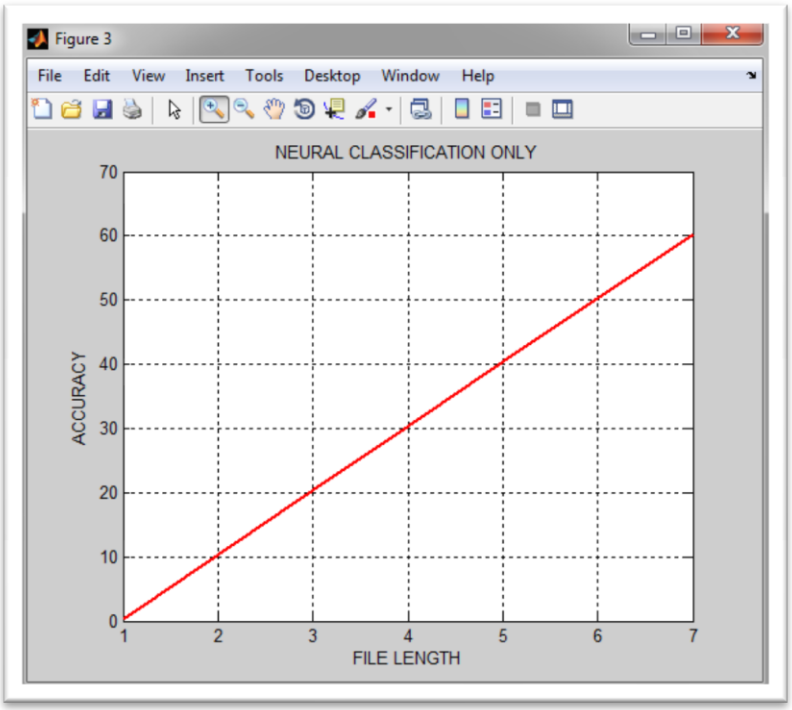

Figure 4.11 Classification accuracy graph of neural only.

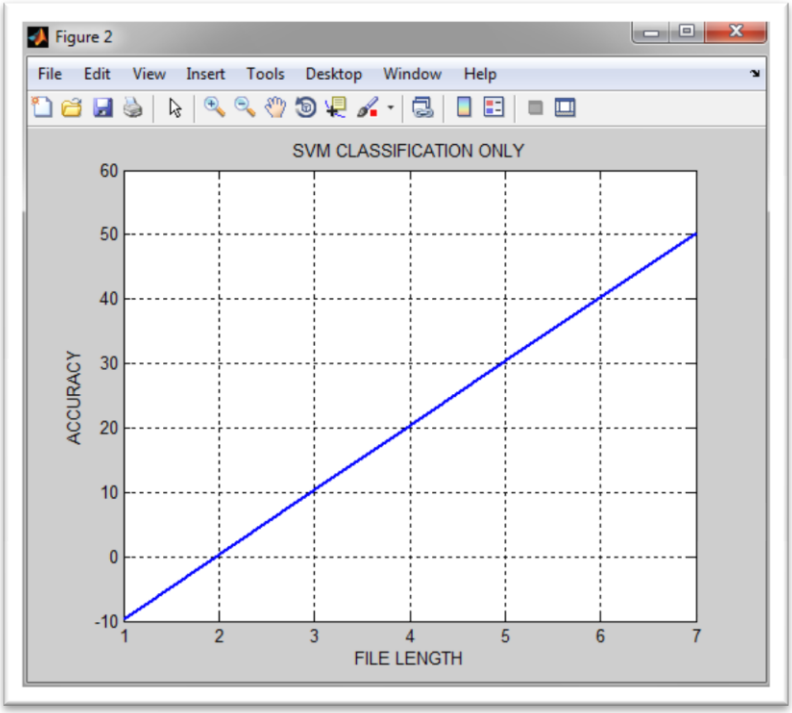

Figure 4.12 Classification accuracy graph of SVM only.

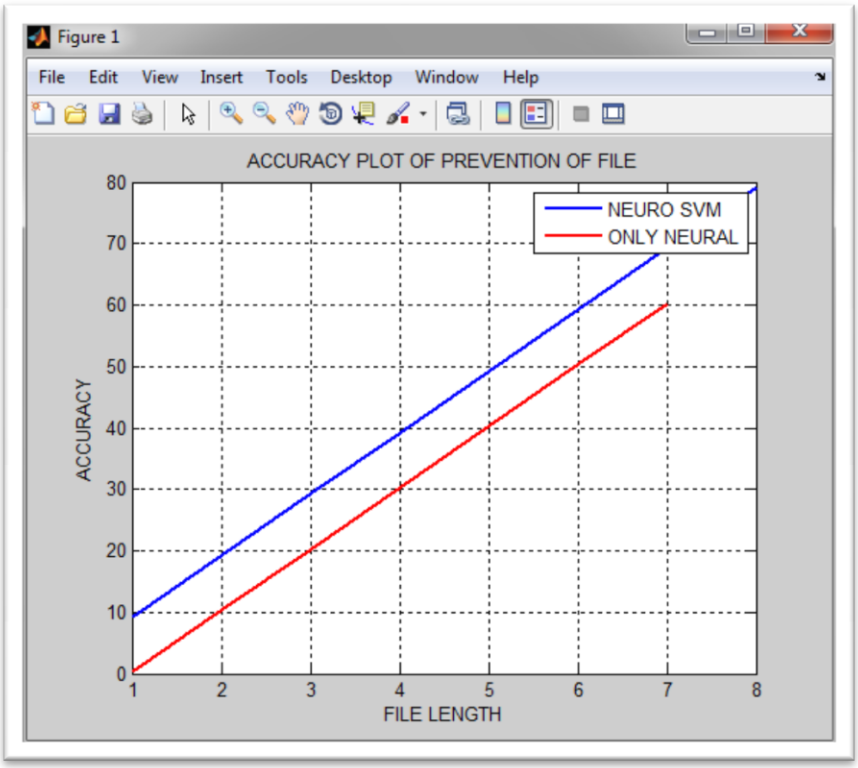

Figure 4.13 Classification accuracy graph of Neural-SVM

Form the figure 4.11, figure 4.12 and figure 4.13 it is clear that the accuracy of neural is almost 60%. Accuracy for SVM is little less than 50%. And the accuracy of Neural-SVM is 79%. This is in the previous case it is 62%, 52% and 76% respectively. So again we have a classification accuracy of almost 20%.

Percentage Accuracy will be calculated as:

Ratio of total number of rows processed of excels spreadsheet to the detected rows of uploaded text file*100.

Chapter 5

Conclusion and Future Scope

Many classifiers can be combined to give better results.. As we discuss text classification with existing technique SVM, Naïve Bayes, and Neural. SVM perform well in case of number system and other dataset. But for online text classification there is performance of neural is better than other existing algorithms. The combination of Neural and SVM performs better from both the previous classifiers. The current research work opens a lot of doors for the future research workers. It is clear from the experimental results that the combination of Neural and SVM improves performance which may improve much more as compared to existing. So for this we can choose Neural and SVM to classify online data like news. In this we find conclusion that SVM is most common useful technique from existing supervised learning other techniques but for online news classification neural is better. The future scope for other researchers is that the can combine 2 or 3 classifiers to get better results. Another is Hierarchal neural network can also be combined with other classifiers like SVM and Naives bayes etc. The current system does not signify any mixed category data and future researchers can upgraded to BFO (Bacterial Foraging Optimization) which is an optimization algorithm but now days researchers are using it for classification.

References

[1] Selamat, Ali, and Sigeru Omatu, "Web News Classification Using Neural Networks Based on PCA", Information Sciences 158; pp. 69-88, 2004.

[2] Zhang, Miao, and De-xian Zhang, "Trained SVMs based rules extraction method for text classification" IT in Medicine and Education, ITME. International Symposium. IEEE 2008.

[3] Amayri Ola, and Nizar Bouguila. "Online spam filtering using support vector machines", Computers and Communications,IEEE Symposium on IEEE 2009.

[4] Meijuan Gao, Jingwen Tian, Shiru Zhou, "Research of Web Classification Mining Based on Classify Support Vector Machine", International Colloquium on Computing, Communication, Control, and Management IEEE CCCM 2009.

[5] Chu Lili Wang Zhuo, "The Algorithm of Text Classification Based on Rough Set and Support Vector Machine", IEEE 2010.

[6] Donghui, Chen, "A new text categorization method based on HMM and SVM."2nd International Conference. In Computer Engineering and Technology (ICCET) Vol.7. 2010

[7] Liu Zhijie, Xueqiang Lv, Kun Liu, Shuicai Shi ,Study on SVM compared with the other text Classification methods." Second IEEE, International Workshop on Education Technology and Computer Science (ETCS), Vol.1. 2010

[8] Wei Zhao and Yafei Wang. "A New Feature Selection Algorithm in Text Categorization", International Symposium on Computer, Communication, Control and Automation, 2010.

[9] Korde, Vandana, and C. Namrata Mahender: "Text Classification and Classifiers: A Survey." International Journal of Artificial Intelligence & Applications (IJAIA), 2012, pp-85-99.

[10] Pushpa, M., and K. Nirmala. "Text Categorization Using Activation Based Term Set" International Journal of Computer Science Issues, Vol. 9, No.3, July 2012.

[11] Mofleh Al-diabat, "Arabic Text Categorization Using Classification Rule Mining", Applied Mathematical Sciences, Vol. 6, pp 4033 – 4046, 2012.

[12] Dasari, Bhavani. "Text Categorization and Machine Learning Methods" Current State of the Art." Global Journal of Computer Science and Technology, Vol.12 Versions 1.0, 2012.

[13] Pratiksha Y. Pawar and S. H. Gawande "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.

[14] Megha Dawar1 and Dr. Aruna Tiwari, "Fast fuzzy feature clustering for Text Classification" in Computer Science & Information Technology Computer Science Conference Proceedings (CS & IT-CSCP), 2012, pp. 167–172

[15] Bhumika1, Prof Sukhjit Singh Sehra2, Prof Anand Nayyar3, "A Review Paper On Algorithms used for Text Classification", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol. 2 Issue 3, March 2013

[16] Harneet Kaur, Dr. Kiran Jyoti, "Design and Implementation of Hybrid Algorithm for e-news Classification", International Journal of Computers & Technology, Dec 2013.

[17] Gayathri, K., Marimuthu," A Text Document Pre-Processing with the KNN for Classification Using the SVM" Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013), pp 453-457, 2013.

[18] Inoshika Dilrukshi, Kasun De Zoysa, "Twitter News Classification: Theoretical and Practical comparison of SVM against Naive Bayes Algorithms", in International Conference on Advances in ICT for Emerging Regions, December, pp 278, 2013.

[19] http://en.wikipedia.org/wiki/Decision_trees#mediaview.jpg (Decision trees)

[20] www.google.com/images/genetic algorithm.jpg

[21] http://users.ics.aalto.fi/harri/thesis/valpola_thesis/node34.html

[22] www.wikipedia.org. (LSI and NLP)

[23] Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, Von-Wun Soo, "Towards Text-based Category Detection: A Survey and Possible Improvements" International Conference on Information Management and Engineering, 2009.

[24] Jones K. S. "A statistical interpretation of term specificity and its application in retrieval" Journal of Documentation, Vol. 28, No. 1, pp. 11-21, 1972

[25] Deerwester S., Dumais S. T., Landauer T. K., Furnas G.W., and Harshman R. "Indexing by latent Semantic Analysis" Journal of American Society of Information Science, 41(6), pp. 391-407, 1990

[26] Zhang W., Yoshida T., and Tang X. Text classification using multi-word Features. In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524, 2007

[27] Changuel S., Labroche N., and Bouchon-Meunier B. “Automatic web pages Author extraction”, Springer-Verlag Berlin Heidelberg, LNAI 5822 pp.300-311, 2009

[28] Zhihang Chen, Chengwen Ni, Murphey, Y.L. International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, July, pp 1054 1060, 2006

[29] Joachims, T, “Text Categorization with Support Vector Machines: Learning With Many Relevant Features Categorization with Class-Based and Corpus-Based Keyword Selection” on European Conference on Machine Learning (ECML’98), 1998