

**REVIEW 1**

**A two-stage feature selection method for text categorization**

BY

**VRUSHIN PARMAR (18BIT0107)**

# AASHISH RAJ (18BIT0113)

**ABSTRACT:**

Content classification or text categorization is used to arrange records in an appropriate way. It arranges any sort of record in different subjects to which it relates and afterward it can demonstrate the primary class wherein it falls. In this procedure, a few issues emerge like addition of noise along with the data. Therefore, feature selection is often used in text categorization to reduce the dimensionality of the feature space and to improve performance. A two phase feature selection and feature extraction is used to improve the performance of content characterization. First phase includes ranking the documents on the basis of their importance of classification (method used will be Information Gain). In the second phase, dimension reduction is carried out on the terms which are ranked in decreasing order of importance. Genetic Algorithm and Principal Component Analysis are applied independently to those terms. So during the process of content categorization, feature selection and extraction methods are only applied to the terms which have high significance; thereby, the computational time and complication of categorization is minimized.

Datasets used are:‐ 1. Reuters‐21,578 and 2. Classic3 datasets

**INTRODUCTION:**

The quantity of content archives in computerized arrangement is continuously expanding and message classification turns into the key innovation to compose content information. Content Categorization or text categorization is the action of naming natural language texts with related categories from a predefined set. A major issue of this process is the high dimensionality of the feature space. This issue may cause the computational complication of ML methods used in text categorization to be increased and may bring about inefficiency in the results. To solve this problem, two techniques are used: feature extraction and feature selection.

Some feature extraction methods have been successfully used such as principal component analysis (PCA) , latent semantic indexing, clustering methods, etc. Out of all the many methods, PCA has got a bit of a more attention. It is a strategy for examination which includes obtaining a linear combination of set of variables that has maximum variance and removing its impact. Feature selection is procedure that chooses a subset from the original set as per some criteria of feature significance. Many of its methods are used for text content classification. But among them, Information Gain proved to be the most effective one. In addition to these, several algorithms inspired from real life such as genetic algorithm (GA) and ant colony optimization algorithm have been used.

**LITERATURE REVIEW:-**

Aghdam et al.[1], proposed an algorithm to deal with feature selections from various kinds of texts. This algorithm is called ‘Ant Colony Optimisation’ and it is based on the methods which ants use in real life while searching for food sources. The author mentions the concept on which the algorithm is based, that is how the ants search for the shortest paths in order to reach their food source. The author compares the proposed algorithm with some already existing algorithms like, information gain, chi-square and genetic algorithm, and compares them with the proposed algorithm which showed the results as the proposed algorithm was having a higher accuracy value (Precision= 77.1343, Recall= 79.7546) on average. The dataset used is the Reuters21578. It can be concluded that this algorithm can provide a very optimal and efficient solution with context to the given problem.

Chang et al.[2], proposed a method which deals with the classifying texts having multiple categories or labels automatically. Various methods were used such as mutual information term selection method, weighted indexing technique and a category-sensitive refinement method which was proposed by them. In the proposed method, Reuters-21578 ModeApte` Split Text Collection was used for training a linear classifier. The author has compared the proposed methods with some existing methods of that time like, Rocchio’s method, K-nearest neighbours (KNN), regression model, Naive Bayes and Bayesian nets, decision tree, decision rules, etc and it showed that the proposed method uses less time for training the classifiers and less classifiers are required to speedily classify the given text. On experimentation, the micro-averaged BEP of the top 10 categories was 87.8 and micro-averaged BEP of all categories was 81.2 for the proposed method, which is quite higher than the other existing methods of that time.

Thomas et al.[3], proposed a semi-supervised method for text clustering. The paper deals with text data only. The focus is on the text classification under the domain of text mining. There are two phases in it:- a training phase and classification phase. Training involves forming clusters from the labelled text present in the data set (Reuters - 21578) and categorizing them according to the text labels. Then, in the classification phase, a new unlabelled text is taken and its similarity is measured (using SMTP)with the centroid of those text clusters and classified accordingly. The use of SMTP provides better accuracy in similarity measurement of unlabelled text. The main focus is on text classification, therefore no focus is made on dimensionality reduction of the text data; this is something to work on.

This research [4] is a survey based report on various machine learning algorithms used in text classification and its different aspects. The author has mentioned two approaches for text classification, namely:- rule based approach which is done manually based on some defined rules, and machine learning approach which is automated text classification approach based on learning from example text/documents. But of them, the later one is quicker with higher accuracy. The author has analysed seven different ML approaches to the problem by different author. The report gives a comparative study of these methods with some external ones.

The research by Lin [5] is an effort towards the energy efficiency aspect of text classification process. Several common classifiers are used and their energy cost is measured using three types of datasets:- 20 Newsgroups dataset by Ken Lang, Reuters-21578, Reuter Ltd. RCV1. So, firstly the author has used classifiers like Naive Bayes, SVM and Perceptron as study objects and then the accuracy and energy cost of those classifiers are measured and compared. Another important thing used is parallelization to reduce energy cost, but it’ll decrease the average power of CPU. Finally it was obtained that the parallel version of Naive Bayes can achieve a high accuracy and is competitive to SVM and Perceptron and it is simple too. It is also an efficient one.

Sabbah et. al [6], has proposed a new frequency based term weighting scheme for accurate text classification which has outperformed other benchmarked schemes significantly. The author has proposed four weighting schemes namely; mTF, mTFIDF, TFmIDF, and mTFmIDF which will take the amount of missing terms into account calculating the weight of existing terms. Datasets like Reuters-21578 R8, 20Newsgroups, and WebKb were used along with few classifiers:- SVM, KNN, NB, and ELM. Finally the result says that the highest performance goes for a SVM classifier with a micro-average F1 classification performance value of 97%. Also mTF, mTFIDF, and mTFmIDF were considered to be the highest performance schemes on analyzing the results.

Thorsten Joachims [7] in 2005 had proposed the idea of using SVM in categorizing text data. He had shown that how exactly SVM can be beneficial to our current problem. Experiments were performed comparing the efficiency of SVM with other methods/classifiers like Bayes, Rocchio, C4.5 and k-NN . The datasets used were:- “ModApte" split of the Reuters-21578 dataset by David Lewis and Ohsumed corpus by William Hersh. The results of the following experiment highlighted the following properties of text:-high dimensional feature spaces, few irrelevant features and sparse instance vectors; and in all these aspects SVM outperformed the remaining.

Mujtaba, et. al [8], has presented a comparative study in identifying the cause of death (CoD) of any person using techniques from text classification. Data for 8 different CoD was obtained from a forensic autopsy reports collected from a few hospitals. Various schemes of text classification such as feature extraction, term weighting, feature reduction, etc were used. Finally the classification model obtained was evaluated and based on 70 dataset instances its overall accuracy was 78.25%, precision = 0.781, recall = 0.783, F-measure = 0.782. But the final model was not good enough for new real time application.

Desmet et al. [9], has given us a very good application of text classification techniques which is suicide prevention. The work is based on forums in the language – Dutch. Genetic algorithms were used in order to optimize the model in a better way of feature selection and hyperparameter optimization. They have used keywork filtering along with some machine learning algorithms for obtaining a good precision-recall value. They have focused the experiment on two kind of tasks:- detecting suicide related posts (in Dutch) and severe, high risk content on internet. The results show high precision and minimal noise. Thus, we can get such system; but the only drawback is that it is for Dutch language only.

Shafiabady, et. al [10], came up with an unsupervised clustering approach for training the Support Vector Machines used in text classification. Training a classifier manually is a very tedious job, but if we can use some unsupervised schemes the clustering part becomes easier. A technique is proposed which uses certain unsupervised methods such as self-organizing maps (SOM) and correlation coefficient(CorrCoef) in order to form clusters for unlabelled text data and then use this data to train SVM. This method also eliminates the problem of dimensionality which is referred as Curse of Dimensionality (COD). The author has experimented with 3 methods to achieve good results. Of all the three methods, the 3rd method in which SVM classifier is trained using “all-to-all” Correlation Coefficient (CorrCoef) produces the best result i.e. highest accuracy. Three datasets are used which are Reuters, Webkb and 20 Newsgroups and the accuracy achieved for them in method 3 are 96.98, 94.73 and 99.72 respectively. Thus, this approach can be used in places where expert decision is not available such as pipeline defect prediction or when the clustering task is too much tedious.

Rehman et al. [11], has obtained a new feature ranking method for selection of most relevant terms while classifying text. Selection of most relevant terms is one of the important part of text classification. The new method used by them is called max-min ratio (MMR). When compared with some of the metrics like balanced accuracy measure, information gain, chisquared, Poisson ratio, Gini index, odds ratio, distinguishing feature selector, and normalized difference measure, this new method proved to be more efficient. Six data sets were used for comparison namely WebACE (WAP, K1a, K1b), Reuters (RE0, RE1), and 20 Newsgroups and classifiers used are multinomial naive Bayes (MNB) and support vector machines (SVM).

The research [12] is a survey done by the author Krina Vasa which deals with some statistical and machine learning approaches that can be applied in text classification. The author has made a survey report for the methods like knearest neighbors, support vector machine, naive Bayesian method, decision tree, rule based classification and neural network. The conclusion of the survey was found as which method provides better performance. So, according to author, The hybrid approach of linear SVM and k-NN provides better accuracy but with the use of kernel function gives better performance than linear SVM.

Viegas et. al [13], has made an attempt to utilize some lazy semi-naive Bayesian strategies effectively and efficiently for text classification. It has been examined whether a correct mix of some alternative NB(Naive Bayes) learning models with distinct feature weighting methods can enhance the efficiency of NB in ADC assignments and then comparing the results with several other supervised algorithms like Nearest-Neighbour classifiers, Support Vector Machines, boosting and some other Bayes algorithms. The experiments findings indicate that a correct mix of learning paradigms and weighting approaches results in several datasets that are similar and even superior to SVMs at reduced cost efficiency. This strategies requires high computational time so they have used GPU parallelization here.

Faraz in his research [14] has presented to us some of the mathematical notations and graphical representations for descriptions of Automatic Text Classification (ATC). A Text Mining Model is also developed which will help to facilitate the design and development of algorithms for Text Categorization (TC) and Automatic Text Classification (ATC) and thus it would enhance the performance of software based on text mining.

Dogan et al. [15], has presented an improved version of inverse gravity method of weighting terms in which two schemes have been highlighted namely SQRT\_TF-IGMimp and TF-IGMimp. The performances of these schemes are then compared with some other standard methods like TF-IDFICSDF, TF-IDF, TF-IDF-ICF, TF-PB, TF-RF, TF-IGMimp, TF-IGM, SQRT\_TF-IGMimp and SQRT\_TF-IGM. Classifiers used are SVM, KNN and NN, and datasets are Reuters-21578, 20 Mini Newsgroups and 20 Newsgroups. The results showed that SQRT\_TF-IGMimp is better than all the other schemes in most cases while TF-IGMimp proved to be better than standard TF-IGM.

Lam et al. [16], has proposed four methods for dimensionality reduction based on artificial neural networks and compared each in terms of precision and recall. A three layer feed-forward neural network was trained and then tested by backpropagation which resulted in a good performance as measured by precision and recall. The techniques used are:- (i)The document frequency (DF) method, (ii) Term occurrence frequency (TF) and the inverse document frequency (IDF) method, (iii) category frequency (CF) and document frequency (DF) method, (iv) method of principal component analysis (PCA). Out of these four methods, PCA proved to be the most reliable one with an reduction rate of 98.9%.

Damerau et al. [17], has presented us the outcomes for automated content classification the accumulation of Reuters-810000 news stories. The author ahs divided the data of into groups of every month and given them an initial standard. For training data, they used two classifiers namely decision tress and linear classifier. The experiment result showed that the linear model is not much feasible for recency effect. Also, it requires less data for training purposes and it is more accurate than the decision trees. However if the categories are less frequent in a data then, decision trees proves to be more accurate.

ElAlami, in his research [18] has mentioned a new feature subset choosing algorithm which with the help of genetic algorithm filters the features which are required to obtain. This algorithm doesn’t modify any training results and neither does it depend on any of the ANN networks to work properly, instead it is only dependent on input features of the training network. The role of genetic algorithm in is to optimize the features which relates to the output function of each class. Later after training this network, this algorithm is applied on Monk1’s and Car Evaluation’s database which shows that the dimensionality of those two databases is reduced by 50% and 33.33% respectively. The results of this experiment is checked against other data mining techniques for the same dataset and this new algorithm proves to be more stable.

Guo, et. al, in his research [19] have proposed a very unique method for weighting the terms in text categorization. The normal term weighting schemes provide only one weight to one term/word, even if it occurs under different labels. The author addresses it as an unreasonable thing to do. So the main idea used by the author is to provide multiple weights to a term so that each of the terms shows its importance in the context. Each weight assigned shows a different class for them. Thus each of these weighted word are then connected as a multi-channel picture which is contribution to a multi-channel CNN model to actualize characterization. The results of this experiment are then compared with other CNN based approaches under the same dataset and this new method proved to be better in most cases.

The research by Ferre [20] is just a comparative study on different method for choosing components for PCA. They made a report on these various methods of how they work, why they failed or why they give very low accuracy. After proper viewing they concluded that there was no perfect solution for the problem of dimensionality reduction so we can just use the most accurate one in these cases. The goal of this paper was to make a descriptive and predictive analysis which was achieved then.

Kępa, et. al [21], has obtained a new classifier approach for text documents. It is a two stage classifier which comprises of kth nearest neighbours(kNN) and SVM classifiers. In their new classifier they have made use of a new method called ‘one-vs-near’ which is an extension of ‘one-vs-all’. In this method, the first classification stage is done by kNN classifier and the next or last stage is done for accurate classification of the trained data. The initial tests were conducted with small sets of data while finally large datasets were used while testing.

Fragoso et al. [22], proposed a new method for filtering the features while performing feature selection, thus to have an efficient vector size for feature selection automatically. The name of this method is Automatic Feature Subsets Analyzer (AFSA) which is an extension of Class-dependent Maximum Features per Document (cMFDR) method. This new method consumes less time in finding the best number of features than the old one. Also, AFSA doesn’t need much training data to produce correct results than that of cMFDR. AFSA selects the similar number of features from the best output of the previous one (cMFDR). Experimental results with datasets namely WebKB, Reuters, 20 Newsgroup, TDT2 proved to provide better or similar results as of that in cMFDR method.

Jiang et al. [23], has proposed a new fuzzy approach for dimensionality reduction of features space. The method automatically generates clusters of words in a feature vector of a text data, based on a similarity test. Then, it extracts one feature from each of the clusters and use it further. As per the algorithm, the user is not required to tell the number of extracted features in advance, and the trial-and-error approach to find those extracted features can also be avoided. Experimental results say that this method be fast and has a good accuracy in obtaining the extracted features.

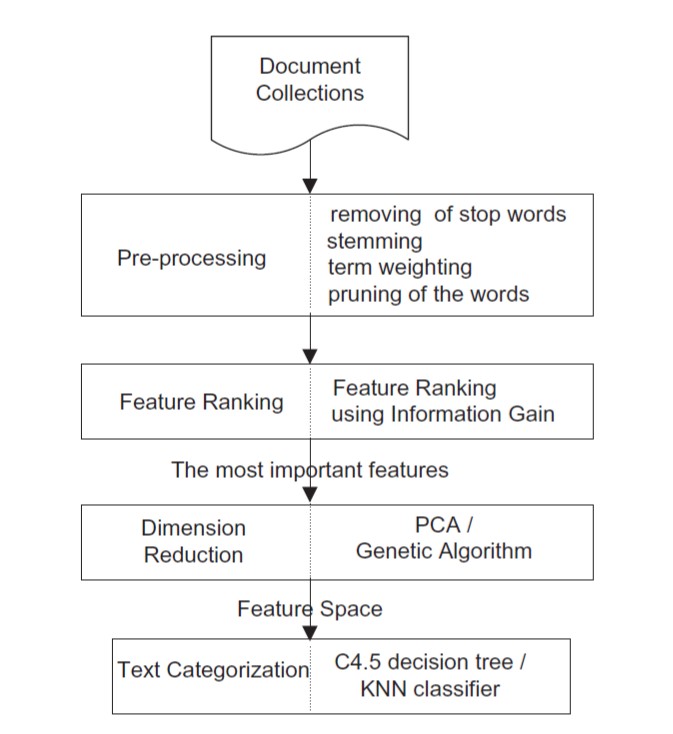
Salles et. al, in his work [24] has pointed out an important factor affecting the old and new algorithms developed for text data classification, which is ‘temporal factor’. Many algorithms have been developed over time, but most of them are developed assuming that data doesn’t change over time. So the author in one of their previous works has shown us some of the adverse effects due to three main temporal effects. In order to minimize these effects on data, a temporal weighting function (TWF) is incorporated with few Automatic Document Classification (ADC) algorithms which they call temporally-aware classifiers, namely Rocchio, KNN and Naive Bayes. These proposed method is evaluated with three real-world datasets that suffer with temporal effects, namely ACM DL, MEDLINE and AG-NEWS. The results tells that temporally aware classifiers significantly improve the results over the traditional ones, with gains up to 17%.

Uysal in his work [25] has proposed an improved global scheme version for feature selection. This scheme is same at all stages except the last stage of a common feature selection scheme is modified. The key aspect behind this method is to equally represent each class in the feature vector. Three distinct datasets were used namely, Reuters, WebKB, Classic3. This method has shown some improvement in performance in terms of Micro-F1 and Macro-F1 metrics.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sr. no. | Author/s and  Year of  Publication | | Dataset | Problem/Objective | Method/s used | Accuracy | Limitations |
| 1. | Aghdam, 2008 | et. al, | Reuters-21578 dataset | Features selection  from texts | Ant Colony Optimization | Precision =  77.1343    Recall =  79.7546 | The parameter values used for testing were determined based on their initial experiments. It is not certain that these vales are optimal. Thus, optimization of parameters can be done. |
| 2. | Chang, 2008 | et. al, | Reuters-21578  ModeApte`  Split Text  Collection | Categorizing text with multiple labels | Mutual information term selection method, weighted indexing technique, categorysensitive refinement method | Microaveraged BEP = 81.2 |  |
| 3. | Thomas, 2016 | et. al, | Reuters-21578 | Classifying text in an efficient manner | Clustering,  Similarity Measure for Text Processing (SMTP) for better accuracy | For similarity  measure –  SMTP, = 186 | No proper dimensionality reduction techniques used.  Thus, more execution time and less varieties of documents can be handled.  It can only handle text data. |
| 4. | Padmavathi.S,  Dr. M.  Chidambaram,  2018 | |  | Survey on different machine learning algorithms used for  classifying text | Boost.SH algorithms, AFSA, inconsistency detection techniques, i-vector method, etc |  |  |
| 5. | Lin, 2015 | | 20 newsgroups  Dataset by Ken  Lang, Reuters-  21578,  Reuters, Ltd. RCV1 | Obtaining an energy efficient way for text  classification | Naive Bayes, SVM and  Perceptron as study objects |  | The only focus is energy efficiency, which is important but other points should also be kept in mind.  Research done only for the three classifiers, few more classifiers like Decision trees could be helpful in the research.  Computation could be moved to GPU for better energy usage. |
| 6. | Sabbah, et. al, 2017 | | Reuters-21578, 20Newsgroups, and WebKB | Method for weighting terms in text  classification | Four frequency-based term weighting schemes:- mTF, mTFIDF, TFmIDF, and mTFmIDF  With SVM classifier | Microaverage F1  = 97 | The method can be used to various feature extraction methods like PCA, and help in solving specific problems.  Better performance could be obtained with the help of some deep learning algorithms and optimization based techniques. |
| 7. | Joachims, 2005 | | “ModApte" split of the Reuters-21578 dataset,  Ohsumed corpus by  William Hersh | Text Categorization | Support Vector Machies (SVM) | Microaverage =  86 | Use of SVM has its own disadvantage like:-  Long training time is required for very large datasets.  Difficulty in choice of kernel.  Highly complex algorithm and extensive |
| 8. | Mujtaba, et. al, 2018 | | Autopsy reports from  University  Malaya  Medical  Hospital, Kuala  Lumpur,  Malaysia | Predicting the cause of death | Various methods of text  classification | Precision = 0.781  Recall =  0.783  F-measure =  0.782 | The prediction accuracy of this model is not good enough for real time deployment.  It can only classify upto eight different cause of deaths.  It can only give cause of death level related to ICD-10. Thus, it cannot give the diagnosis upto a greater detail. |
| 9. | Desmet, Hoste, 2018 | | Dutchlanguage forum and blog messages  posted on  Netlog | Suicide prevention by identifying related  texts | Genetic algorithms |  | It only works for labelled data, so we have to be dependent on that. The data presented and used is specific to Dutch. No multi-lingual support. Reduction of linguistic noise could be done. Text variation could be reduced.  Improvement in lexical recall is required. |
| 10. | Shafiabady, et. al, 2016 | | 20  Newsgroups,  Reuters-  21578(R8) and  WebKB | Training of Support Vector Machine for  text classification | Self-Organizing Maps (SOM), Correlation Coefficient  (CorrCoef) | For method-3 in dataset:-  Reuters =  96.98  Webkb =  94.73  20  Newsgroups  =99.72 |  |
| 11. | Rehman, et. al, 2018 | | WebACE (WAP, K1a, K1b),  Reuters (RE0, RE1), and 20  Newsgroups | Finding most relevant terms in a text data | Max-Min Ratio (MMR), multinomial naive Bayes (MNB) and support vector machines (SVM) classifiers | Macro F1  measure =  76.2  Micro F1 measure =  74.4 | There is a need to develop effective local or class-wise feature selection algorithms.  Work needs to be done on weighting scheme because it is an integral part of the process, no relevant works should be removed from the document. |
| 12. | Vasa, 2016 | |  | Survey on statistical and machine learning approaches of text  classification | k-nearest neighbours, support vector machine, naive Bayesian method, decision tree, rule based classification, neural network |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13. | Viegas, et. al, 2018 | | 20 Newsgroups, Four Universities, Reuters, ACM Digital Library, RCV1, AGNews, Medline and Yahoo datasets | | Classifying text in an effective and efficient manner | | Lazy Super Parent Tree Augmented Naive Bayes (LSPTAN) method | | MacF1 and MicF1 for:--  20 Newsgroups = 90.58 ±1.16,90.77 ±1.16  Four Universities = 61.10 ±2.17, 71.60 ±1.84  Reuters = 37.71 ±1.86 ,66.74 ±1.10 ACM Digital Library = 64.32 ±1.67 ,75.93 ±0.71 RCV1 = 55.55 ±0.67,73.25 ±0.28 AGNews = 61.56 ±0.03 ,68.89 ±0.09 Medline = 71.69 ±0.57 ,84.61 ±0.03 Yahoo dataset = 66.30 ±0.11,66.62 ±0.11 | | High computational costs such as time.  In large Automatic Document Classification (ADC) datasets, its performance is minimized. | | |
| 14. | | Faraz, 2015 | |  | | Text mining with a mathematical approach | | Mathematical notations and graphical modelling techniques | |  | | Mathematical models have limited scope in data mining |
| 15. | | Dogan, et. al, 2019 | | Reuters-21578, 20 Mini Newsgroups and 20 Newsgroups | | An enhanced term weighting method for text classification | | Inverse gravity moment formula(SQRT\_TF-IGMimp and TF-IGMimp);  \*imp means improved version | | For Reuters-21578, the Macro-F1 using SVM classifier is 96.57 (highest) | |  |
| 16. | | Lam, et. al, 1999 | | A subset of the Reuters-22173 test collection | | Neural networks based feature reduction for categorizing text data | | DF Method, TFxIDF Method, CF-DF Method, Principal Component Analysis | | Reduction rate of 98.9% | | PCA works on the basis of certain assumptions. It may give inaccurate results for a large variety of data. Also, it is a linear method. |
| 17. | | Damerau,et. al, 2004 | | Reuters-810000 collection | | Text categorization | | Decision tree classifier and linear classifier | |  | | Amount and recency of data used in training is dependent on the model |
| 18. | | ElAlami, 2009 | | Monk1’s database and Car Evaluation database | | Choosing a subset of features from a feature set | | Genetic algorithm | | For Monk1’s database - 50%,  Car Evaluation database – 33.33% | | Testing is done with a limited data set type. |
| 19. | | Guo, et. al, 2019 | | Movie review sentence polarity dataset v1.0, subjectivity data set, customer reviews  of 14 products from Amazon, MPQA, TREC | | Improved way for term weighting in text classification | | A multi-channel Text-CNN model | | Accuracy for,  MR = 86.6,  Subj = 95.6,  CR = 87.5, MPQA = 93.0, TREC = 91.9 | | Size of the datasets used is not too large here.  There are various word installing networks in our circumstance, it is thus hard to concentrate or posture how each inserting framework is identified with one another |
| 20. | | Ferre, 1995 | |  | | Comparing methods for selecting components for PCA | | Different methods of component selection in PCA | |  | | There is no proper or totally correct solution for solving the reducing the dimensions during PCA |
| 21. | | Kępa, et. al, 2015 | | Wikipedia | | Classification of text documents in two stages | | A new method named one-vs-near | | Precision = 59.40%  Recall = 33.34% | |  |
| 22. | | Fragoso, et. al, 2016 | | WebKB, Reuters, 20 Newsgroup, TDT2 | | Determination of vector size for feature selection in text categorization | | Automatic Feature Subsets Analyzer (AFSA) | | For TDT2 dataset, Micro- F1 value = 96.68 | |  |
| 23. | | Jiang, et. al, 2011 | | 20 Newsgroups Dataset, Reuters Corpus Volume 1 (RCV1) Data Set, Cade12 Data | | Reduce dimensionality of feature space for text classification | | Fuzzy self-constructing feature clustering (FFC) algorithm | | For Cade12 dataset,  Micro averaged accuracy = 93.55 | |  |
| 24. | | Salles, et. al, 2017 | | ACM DL, MEDLINE and AG-NEWS | | Effiecient time-based text classification | | A temporal weighting function (TWF) incorporated with classifiers namely Rocchio, KNN and Naive Bayes | |  | | This may affect the overall speed and performance. |
| 25. | | Uysal, 2015 | | Reuters, WebKB, Classic3 datasets | | A global scheme for features selection in documents classification | | Improved global feature selection scheme (IGFSS) | | Micro-F1 scores (%) for Classic3 dataset = 98.973 | |  |

**METHODOLOGY:**



The dataset we are using is Reuters-21,578 data set. From this data set we are selecting six categories and this includes 8158 documents. The chosen categories along with number of instances are displayed in Table 1.

**Table 1**

|  |  |
| --- | --- |
| **Category Name** | **No of documents** |
| Earn | 3965 |
| Acquisition | 2369 |
| Money-fx | 719 |
| Crude | 580 |
| Grain | 583 |
| Trade | 486 |

(A) Pre-processing:

In pre- processing stage we will remove the stop words (‘a’,”and”,”the”,etc.) then we will go for stemming. In stemming we will found the root forms of the word ( like for “computer”,”computation”,”computes” the root word is comput ).

After this we will go for term weighting and pruning of words( removing words with less frequent features).

We have made use of Natural language Toolkit(NLTK), which helps in this pre-processing part.NLTK - The Natural Language ToolKit is one of the best-known and most-used NLP libraries in the Python ecosystem, useful for all sorts of tasks from tokenization, to stemming, to part of speech tagging, and beyond. It has helped in the process of stop words removal, tokenizing and stemming of data.

(B)Feature ranking:

In feature ranking stage we will rank the terms as per their importance. We will be using Information Gain method for this. The idea of Information gain is based on information theory.

In Information gain, we obtain the amount of information that will be gained by knowing the value of any feature. Now, the data obtained from the pre-processing step will become our training data. This training data is used time and time again to find the information gain, gini impurity, etc. We will now ask a series of questions to the machine that will be useful for feature ranking process; for that first we need to form the questions that are best to ask, this will be decided using information gain of those features.

An important aspect to calculate information gain is gini impurity. When we ask questions, we need to find the right or the best question in order to simplify the process. The best question is the one that reduces our uncertainty the most; and gini impurity helps us quantify how much uncertainty is present. Thereby, Information Gain helps us quantify that much a question reduces that.

(C)Dimension reduction:

After the pre-processing step, even though the terms of extreme importance are obtained already, there still exists the high dimensionality of the feature vector. After the previous step, each term within the text are ranked and then dimensionality reduction is carried out for those terms which are of extreme importance.

For dimensionality reduction we have used the method called Principal Component Analysis(PCA).In PCA algorithm, we reduce the dimension of a multi-dimension data.

We search for axis for which we should get the least overlap of the labels when projected in a single dimension. We use the concept of eigen values and vectors for this, as product of eigen vector for a given matrix and the same matrix represents the matrix with a same direction but with greater magnitude. We will consider those eigen vector having highest eigen value. This will make reduce the chance of overlapping of labels while projecting on a single dimension axis.

(D)Text Categorization:

Finally, we will evaluate the effectiveness of dimension reduction methods on our proposed model experiments conducted using KNN (K-Nearest Neighbour) and C4.5 decision tree.

**SOFTWARE/TOOLS/LANGUAGE:**

SOFTWARE: ANACONDA, SPYDER, JUPYTER NOTEBOOK

LANGUAGE: PYTHON

**RESULT:**

SOF

**CONCLUSION:**

In the current study, a two-stage feature selection and feature extraction is used. We reduce the high dimensionality of a feature space, remove redundant and irrelevant features from it and thus improve the performance of text categorization. In the initial stage, we will rank the features of each terms obtained from the pre-processed text data. The ranking is done using the method of Information Gain and decision trees. In the latter step, we apply Principal Component Analysis to those terms having more importance and thus reduce the dimensions of the feature vector space. Therefore finally text categorization is carried out on the terms with higher importance and because of that the computational time is greatly reduced.

**REFERENCES:**

1. Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri, “Text feature selection using ant colony optimization”
2. Yu-Chuan Chang, Shyi-Ming Chen, Churn-Jung Liau, “Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method”
3. Anisha Mariam Thomas, Resmipriya M G, “An Efficient Text Classification Scheme Using Clustering”
4. Padmavathi.S, Dr. M. Chidambaram, “A Brief Survey on Text Classification Using Various Machine Learning Techniques”
5. Hao Lin, “Research on Energy-Efficient Text Classification”
6. Thabit Sabbah, Ali Selamat, Md Hafiz Selamat, Fawaz S. Al-Anzi, Enrique Herrera Viedma, Ondrej Krejcar and Hamido Fujita, “Modified Frequency-Based Term Weighting Schemes for Text Classification”
7. Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”
8. Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh, “Prediction of Cause of Death from Forensic Autopsy Reports using Text Classification Techniques: A Comparative Study”
9. Bart Desmet, Véronique Hoste, “Online suicide prevention through optimised text classification”
10. Niusha Shafiabady, L.H. Lee, R. Rajkumar, V.P. Kallimani, Nik Ahmad Akram, Dino Isa, “Using unsupervised clustering approach to train the Support Vector Machine for text classification”
11. Abdur Rehman, Kashif Javed, Haroon A. Babri, Nabeel Asim, “Selection of the Most Relevant Terms Based on a Max-Min Ratio metric for Text Classification”
12. Krina Vasa, “Text Classification through Statistical and Machine Learning Methods: A Survey”
13. Felipe Viegas, Leonardo Rocha, Elaine Resende, Thiago Salles, Wellignton Martins, Mateus Ferreira e Freitas, Marcos André Gonçalves, “Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification”
14. Ahmed Faraz, “An elaboration of text categorization and automatic text classification through mathematical and graphical modelling”
15. Turgut Dogan and Alper Kursat Uysal, “Improved inverse gravity moment term weighting for text classification”
16. SavioL. Y.Lam & Dik Lun Lee, “Feature Reduction for Neural Network Based Text Categorization”
17. Fred J. Damerau, Tong Zhang, Sholom M. Weiss, Nitin Indurkhya, “Text categorization for a comprehensive time-dependent benchmark”
18. M.E. ElAlami, “A filter model for feature subset selection based on genetic algorithm”
19. Bao Guo, Chunxia Zhang, Junmin Liu, Xiaoyi Ma, “Improving text classification with weighted word emb e ddings via a multi-channel TextCNN model”
20. Louis Ferre, “Selection of components in principal component analysis: A comparison of methods”
21. Marcin Kępa and Julian Szymański, “Two Stage SVM and kNN Text Documents Classifier”
22. Rog´erio C. P. Fragoso, Roberto H. W. Pinheiro, George D. C. Cavalcanti, “A method for automatic determination of the feature vector size for text categorization”
23. Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, “A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification”
24. Thiago Salles, Leonardo Rocha, Fernando Mour˜ao, Marcos Gonc¸alves, Felipe Viegas, Wagner Meira Jr., “A Two-Stage Machine Learning Approach for Temporally-Robust Text Classification”
25. Alper Kursat Uysal, “An improved global feature selection scheme for text classification”