ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

# Data Mining Method using Clustering Mechanisms and Feature assortment for efficient content categorization

[1] Sri. E. Ramesh, [2] Dr. B Tarakeswara Rao
[1] Assistant Professor, [3] Professor
[1] Department of CSE,RVR & JC College of Engineering, Guntur, Andhra Pradesh. [2] Department of CSE, Kallam Haranadhareddy Institute of Technology, Guntur, Andhra Pradesh

***Abstract* - Data mining manages the way toward finding data from data. With the wide accessibility of data there are different applications and requirements for data mining. Data can be of any sort specifically message, pictures, recordings and some more. This work concentrates on characterization of content data utilizing a semi-regulated grouping calculation. The principle issue in any data mining assignment is the treatment of gigantic data. Immense dimensionality does not demonstrate more data rather it might incorporate irregularities and commotions. To make the data predictable powerful pre-preparing strategies are finished. Notwithstanding this element choice system chooses valuable highlights and evacuate superfluous ones along these lines making the data significant for mining. A semi managed bunching calculation TESC was utilized as a part of this investigation. It was then changed by actualizing a component determination strategy, record recurrence thresholding by which the tremendous dimensionality issue was tended to. The proposed framework along these lines beats the current strategy. Tests were led on Reuters-21578 which portrayed better execution with lessened time many-sided quality.**

**Index Terms—Classification, Clustering, Data Mining, Feature Selection, Text Mining**

## 1. INTRODUCTION

With the quick development of advances like web the data being created from such different sources additionally expanded. Colossal measure of data is being produced and is overflowing with data. To make these data valuable advances like learning disclosure and data mining rose. Learning disclosure and data mining are firmly related. Learning revelation is the way toward finding data or data from gigantic data while data mining is the use of calculations and techniques for finding data. Since the data age rate is expanding quickly so is the significance of this field. So inquires about are being led to perceive better strategies. The exploration viewpoints in this field incorporate dealing with the data, execution and proficiency of calculations and introduction of data. Data mining utilizes different machine learning strategies like bunching, grouping and relapse for data recovery. This work concentrates on grouping and order of content data. Text data is an accumulation of archives that can be classified by its substance. Each record is made out of terms which might possibly be helpful. Text preparing is a branch that robotizes the way toward controlling the first data as indicated by some particular application areas. It incorporates designing[3], seeking, producing a few examples or results, separating and so forth. In the present situation of colossal and detonating electronic data accessible from different sources it is extremely important to extricate data successfully[8]. This need builds the extent of research in the field of data mining. There are different grouping calculations created because of broad research. In any case, accomplishing characterization[11] impeccably is still not feasible. So the scientists attempt to enhance the execution of existing techniques by actualizing new ideas or some cooperative thoughts.

Text grouping is a dynamic research zone in which orders the content record to the classifications[9] they have a place with. It incorporates two stages viz, grouping or preparing stage and a model creation stage. The contribution to the grouping stage is a report set D = {d1, d2,… dn}, which is named to a classification L $\epsilon$ C . Amid this errand the comparative classifications are recognized and bunched into comparative gatherings. Records inside similar bunches are comparative and are not quite the same as those in another group. Utilizing this grouping or preparing data an arrangement display f is made which maps the archives to their particular classifications. i.e.

$$f : D * C; \; f(d) = L \qquad (1)$$

The characterization display along these lines appoints

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

each record to the suitable name. Therefore the model can be utilized to order other comparative archives by which we can test the proficiency of the model made. In testing the model the records whose names are as of now known are given to the model. It at that point registers a name as the yield of the model. Terms or segments of the records are the key on-screen character[5] in characterization and grouping of data. Since the terms and classes might be corresponded the undertaking of grouping turns into a troublesome errand.

This trouble increments as the measurement of the vector space that speaks to an archive increments. Because of these reasons a widespread grouping model is as yet a unimaginable errand [1].

This work concentrates on the most proficient method to execute an effective technique contrasted with existing ones. From the current study in grouping of content data a semi-administered bunching calculation to be specific TESC(Text order utilizing Semi-regulated Clustering )[2] that beats different existing techniques was picked as a base for this examination. It is an inventive and diverse technique from the current ones. So as to upgrade the execution a component selection[3] strategy consolidated calculation was proposed. The proposed framework performs superior to the current technique by maintaining a strategic distance from the gigantic dimensionality and time many-sided quality issues.

Whatever is left of this paper is composed in the accompanying segment. Area II is a broad survey and investigation of the fundamentals and points of interest of grouping errand and different strategies embraced. Area III depicts the strategies and steps embraced in executing the proposed framework. Segment IV portrays the execution assessment and investigations led. Segment V finishes up the investigation by a short examination of the proposed framework.

## 2. RELATED WORKS

Data mining have shaped a branch of connected computerized reasoning since the 1960s. The different utilization of data mining incorporates neural systems, calculation engineering, dynamic expectation, investigation of framework design, smart frameworks, displaying, learning based frameworks. Distinctive sociology procedures, for example, brain research, intellectual science[13] and human conduct can be utilized as a part of combination with data mining techniques will build execution . Data mining can likewise be incorporated with the cutting edge advances like IoT, Data mining and Big data[4]. It can be valuable in areas that incorporate web based business, ventures, for example, retail, managing an account, broadcast communications, social insurance, open administration region, criminal discovery, transport frameworks, and some more.

Bunching is a machine learning technique embraced in data mining which consolidates the assignment of collection comparative data together. The comparative gatherings accordingly shaped are called groups. There are different kinds of bunching to be specific progressive, dividing, chart based, demonstrate based, lattice based and delicate figuring based strategies. By and large there are two stages to be specific the preparation stage and the testing stage[14]. In preparing the model first calls calculation for preparing the info dataset and makes a characterization demonstrate. In the testing stage it assesses the model made. In view of the info dataset grouping might be ordered as administered and unsupervised. In regulated grouping the data incorporates both the data and the coveted outcomes. In unsupervised grouping the model isn't given the coveted outcomes amid the preparation. In unsupervised models bunching is done in light of factual properties. Notwithstanding these semi-administered bunching calculations are as of late ending up extremely prevalent as a result of the tremendous accessibility of unlabelled data [5].

In semi-administered bunching the named or known records are first grouped and after that unlabelled reports are fitted into the correct marks utilizing some similitude measures. In semi-administered grouping an archive accumulation D is thought to be an accumulation of named DL and in addition unlabelled reports DU ie, D = { DL , DU }. DT is a subset of D which is utilized for preparing the framework demonstrate f. Semi managed bunching calculation finds a parcel C utilizing DT = { , } where DL and DU . The parcel C = {C1, C2 , … Cm} where $1 \leq I \leq m$ Ci = DT and Ci ∩ Cj = ∅. At the point when another known info di is given to the model, it allots the contribution to the proper parcel in light of some likeness measure.

The execution of arrangement calculation is subject to the nature of data source[6]. So pre-preparing is an imperative advance in data mining as it characterizes the important contribution to the framework. It incorporates the accompanying strategies. Tokenization is the procedure by which the tokens or words are distinguished from the archive. The

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

procedure stop-words evacuation recognizes the uncommon characters[12] in the report and are expelled from the contribution to diminish the insignificant data. Standard stop words list are accessible on the web. Stemming is the procedure which diminishes words to an essential type of the word for that it strips the postfix and prefix of the word.

Data mining handles diverse kinds of data including very much organized, semi organized and even unstructured reports. To speak to a pre-handled report there are different procedures like TF-IDF, LSI, multi-words and so forth [7] among which TF-IDF is a proficient strategy and is being utilized with different structures for text classification. The key idea of TF-IDF is that a given archive can be sorted in light of regardless of whether the term is applicable with the name of a given report. In light of the comparability measure between vectors the bunching procedure is finished. A ton of measures have been proposed for processing for similitude measure. The Kullback-Leib ler disparity, Euclidean separation, Manhattan remove, Canberra remove metric, Cosine similarity, Bray-Curtis, Jaccard coefficient, Hamming separation and so on are couple of well known techniques. Among the different models the Euclidean and Cosine measures are the most ordinarily utilized methods[8].

A noteworthy issue in content characterization is the colossal dimensionality of the vector space. It is extremely important to decrease the vector space without trading off the classification undertaking. For this reason different component choice techniques like Document Frequency, Data Ga in, Mutual Data and so forth are used[9]. Report Frequency of a term is the quantity of records in which the term happens. DF thresholding consequently processes the report recurrence for each term and expels the terms recurrence are not as much as a foreordained edge. Data Gain is a machine learning system that considers term goodness standard. It gauges the data got for a classification expectation by knowing the nearness or nonappearance of a term in a record. Common Data is factual dialect demonstrating approach. It thinks about term and classification and considers the nearness and nonattendance of terms. Among the different strategies, Document Frequency Thresholding is one of the most straightforward and productive technique that give preferable exhibitions over different measurable methods[10].

There are different strategies existing to actualize text order and arrangement. Characterization utilizing semi regulated bunching, grouping fusing highlight determination, [11] characterization through consolidated methods[12], gathering methods[13] and so forth are different usage of grouping that are not the same as the conventional techniques. In the current strategy for content arrangement utilizing semi-directed grouping calculations there are different methodologies in particular Semi-Supervised Cluster (SSC) tree technique [14]. The vast majority of the current strategies utilizes k means and its variety for executing the semi-regulated strategy. The calculation TESC is a current and imaginative approach which gives better outcomes and execution in characterization when contrasted with the current ones. The principle entanglement in any order assignment is the colossal dimensionality of the data. The colossal measure of data won't not contain helpful data alone. Different element determination techniques are utilized to choose the imperative highlights that streamline the assignment of arrangement. Among the different component choice techniques an archive recurrence thresholding was incorporated into the calculation TESC to enhance the execution.

## 3. PROCEDURE

In this work the bunching and grouping of content data is the primary region of study which toss in together some e xisting ideas to make the way toward grouping and arrangement more successful. The essential thought is grouping the enormous content reports that have a place with a few names by considering different components. This work concentrates on semi-administered bunching process which makes utilization of named and unlabelled data for grouping archives. The principle issue in grouping of content data is the colossal measurement. To settle this issue an element determination technique is received by which the commotion are expelled. The class of every content bunch is named by the terms it contains. At the point when another unlabelled content shows up the closeness measure of the new report with the bunches distinguished is figured. At that point the data is ordered to the closest mark. In some datasets to enhance the execution an idea of highlight choice is received.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

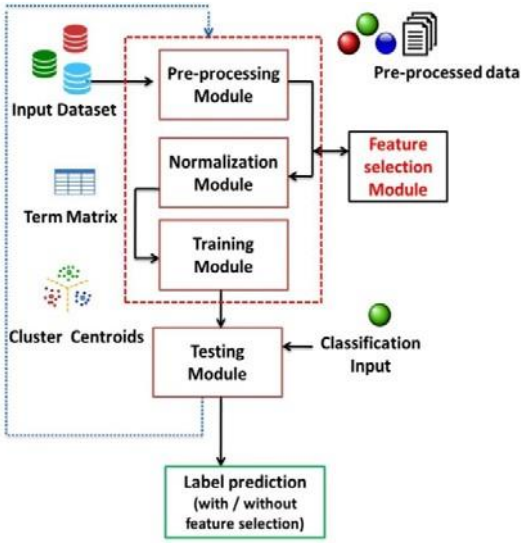


***Fig 1. Proposed System Design***

### *A.System Design*

Text grouping in view of semi-administered bunching is a dynamic territory of research. In this investigation a straightforward and imaginative calculation TESC(Text- classification utilizing Semi-administered Clustering)[2] was broke down and executed. This calculation makes utilization of both marked and unlabelled data for bunching the data. The named archives are first grouped and after that the unlabelled data adjust to the named bunches. The outline of the general framework depends on a basic idea and is as appeared in the Figure 1 which is a point by point portrayal of the proposed demonstrate.

### *B.Preprocessing of Input Dataset*

The as a matter of first importance venture in all data mining system is a preprocessing stage. It is a vital stage in light of the fact that the nature of yield delivered relies upon the nature of info given. The data for an data mining system might be an data gathered from any electronic medium. So there are odds of the data to be deficient, boisterous and conflicting. Consequently the change of the info steady with the every application area preprocessing steps are done[6]. Therefore the preprocessing stage is viewed as an unavoidable stage in all data mining and learning disclosure process. In this investigation the Reuters-21578 [15] dataset was utilized as data.

1). Extracting the archives and marks from the current data collection : Each record has a place with a specific subject. There are 135 points in the Reuter 21578 dataset. For the simplicity of usage the dataset can be diminished or modified preparing set by every application area. From the dataset the reports and the classes to which they have a place are extracted.

2). Stop words expulsion: Stop words will be words which are sifted through earlier or in the wake of handling of characteristic dialect data. It alludes to the regular words in a dialect. For example in English the words the, of, and so on are stop-words. These regular words don't give any detect to grouping and characterization rather it gives extra overhead while handling. So to reduce the complexity factors like computational time and dimensionality it is important to evacuate these undesirable terms. This procedure is called stop-words evacuation. The usually utilized stop words utilized for examine objects are openly accessible. These rundowns were included to a cluster list and the preprocessed reports were cross approved to check the events of stop-words. All the stop-words consequently found was expelled from the record.

3). Tagging: The archive is subjected to a procedure called labeling in which the parts of discourse of the terms in the report are distinguished. There are different strategies for actualizing labeling. A current tagger the Stanford Log-straight Part-Of-Speech Tagger [16] was utilized for this framework. It is a bit of programming that peruses message in some dialect and appoints parts of discourse to each word, for example, thing, verb, descriptor, and so on. The POS tagger is actualized in Java and is accessible on the web.

4). Stemming : Stemming is the way toward decreasing words to their base or root form.
finish and reliable in this way giving a quality contribution to the framework.

### *C. Normalizing the Cleaned Dataset*

Standardization of the dataset is an essential stage in data mining. There are different techniques for dealing with the dataset normalizing as interoperability is a vital viewpoint. All parameters ought to have a similar scale for a relative examination between them. Standardization is such an data rescaling strategy by which the numeric esteems are scaled in the range [0,1]. For this undertaking the Term Frequency-Inverse Document Frequency(TF-IDF) idea is utilized. It is a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

factual measure used to assess how vital a term is to an archive in a corpus. The significance expands relatively to the quantity of t imes a word shows up in the archive yet is counterbalanced by the recurrence of the word in the corpus. The TF-IDF weight is made by two terms to be specific Term Frequency (TF) and the Inverse Document Frequency (IDF). Following definition demonstrates to figure this weight[17]. The Term Frequency measures how as often as possible a term shows up in an archive. In a preparation corpus the archives may fluctuate long and contain wide assortment of terms. A recurrence of term showing up might be more than the length of the records. In this way, the term recurrence is isolated by the report length ie, the aggregate number of terms in the archive considered. The condition for the count of TF of a term t is given in underneath:

$$TF(t) = \frac{\text{(Number of times term t appears in a document)}}{\text{(Total number of terms in the document)}} \quad (2)$$

The Inverse Document Frequency measures how critical a term is. While processing TF, all terms are considered similarly essential. Be that as it may, to separate the terms with more significance noticeable we have to overload the incessant terms wh ile scale up the uncommon ones. The condition for figuring IDF is given underneath :

$$IDF(t) = \log_e \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}} \quad (3)$$

After calculating TF and IDF the overall weight can be obtained as a product of these individually calculated weights

TF-IDFweight = TF(t)*IDF(t) (4)

The output is a matrix containing all terms in all documents and their corresponding TF-IDF value. Using this matrix the similarity measure between vectors are calculated and clustering is done.

### *D. Clustering of the Dataset*

The bunching procedure distinguishes segments from named and unlabelled content. The marked content is utilized for managed learning and utilizing this unlabelled writings adjust the centroids of content parts. A basic calculation which beat Support Vector Machines and back proliferation neural system and produces practically identical execution to Naïve Bayes with EM (Expectation Maximization) with bring down calculation many-sided quality was recognized by writing audit The calculation was named TESC: Text order utilizing Semi-directed Clustering. The bunching procedure embraced in this work incorporates three stages to be specific initialization, grouping and yield.

### *E. Classification*

Order is the way toward testing which decides the proficiency of the bunching strategy received. In this stage the referred to marked archive is given as a testing contribution to the framework. The framework at that point actualizes the grouping of the info utilizing the preparation learning and registers a mark as yield. On the off chance that the name anticipated through the calculation is same as that of the info mark then the strategy executed effectively anticipated the yield. Be that as it may, progressively application space idealize expectation precision is incomprehensible.

### *F. Feature Selection*

To enhance the execution of the current framework the degree for joining an element choice approach was dissected. There are different methodologies like report recurrence thresholding, data increase, common data increase existing. In this work report recurrence thresholding was executed. It is a basic and benchmark approach however performs similarly or superior to other unpredictable and factual techniques.

The archive recurrence gives the quantity of records in which a term or vector show up. Just the terms with recurrence more noteworthy than an edge is taken others are disposed of. By this we can decrease the aggregate number of highlights or vectors. So after the records are pre-handled a report recurrence thresholding process is done to diminish the vocabulary. At that point a standardization stage is done which computes the TF-IDF.

In the proposed strategy we acquire a nearly littler TF-IDF lattice. It isn't just a technique for vocabulary decrease yet in addition concentrates on determination of critical or successive terms for productive arrangement. The system is demonstrated as follows:

Info: Set of named and unlabelled records, df threshold, test data

Yield: Classified reports Procedure:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

1.For all named and unlabelled archives set the underlying record names and bunch hopeful set. Apply archive recurrence thresholding
2.Calculate the recurrence of terms in archives
3.If the estimation of the recurrence is not exactly df threshold esteem
   a. Remove the terms
4.Else
   a. Retain the terms

5.Calculate the TF-IDF angle.

6.Apply the semi-directed grouping calculation
7.Test the records - characterization

## 4 . RESULTS

Various tests were led and an investigation of the execution of the proposed framework was finished

***A. Data Set :*** For investigating the execution of the framework a progression of experiments were led with the Reuters data collection. It was the isolated into two to be specific the preparation data and testing data. The preparation data collection utilized was a blend of few named data (acq,grain and rough) and couple of unlabelled data. After the preparation the unlabelled data was bunched to the known marks. Testing data is the contribution for the arrangement stage.

***B.Experiments :*** The framework was then tried with 30 testing datasets each for every one of the names. Each info was ordered in light of the preparation dataset which included named and unlabelled data. The bunch centroids for the new approaching yield is distinguished in view of the group centroids of the preparation set. Every name acq,grain and rough were renamed to name y1,y2 and y3 separately. The forecasts for 30 contributions for every mark are examined. The forecasts in light of Method 1(TESC) and Method 2(TESC+DF) were acquired broke down.

### *C. Evaluation and Analysis of Results*

•Analysis in view of Correct Predictions : As arrangement of experiments directed on Method 1(TESC) and Method 2(TESC+DF) and the outcomes were broke down as beneath. To dissect the execution of the characterization a perplexity grid strategy was utilized. It is a table that is utilized to depict the execution of an order demonstrate on testing data for which the bona fide qualities are known. The tables I and II demonstrates a perplexity lattice of both the techniques executed. The left most sections shows the genuine marks of the records and best most lines demonstrates the anticipated names. The disarray lattice takes a couple of names <l1, l2> and examinations what number of records from l1 were mistakenly doled out to l2. In Table I and II the classifier recognizes three marks to be specific y1,y 2,y3 yet makes numerous mistakes inside. The perplexity grid can help pinpoint open doors for enhancing the precision of the framework

***TABLE I : CONFUSION MATRIX FOR METHOD 1***

| Labels | y1 | y2 | y3 |
|---|---|---|---|
| y1 | 17 | 5 | 8 |
| y2 | 8 | 16 | 6 |
| y3 | 9 | 6 | 15 |

From Table I for mark y1, 17 archives were accurately named y1 however 13 reports were anticipated wrongly as y2(5) and y3(8). Essentially for y2, 16 records and for y3, 15 reports were accurately anticipated as y2 and y3 separately

***TABLE II : CONFUSION MATRIX FOR METHOD 2***

| Labels | y1 | y2 | y3 |
|---|---|---|---|
| y1 | 18 | 4 | 8 |
| y2 | 6 | 20 | 4 |
| y3 | 9 | 3 | 18 |

From Table II for name y1, 18 archives were accurately named y1 however 12 reports were anticipated wrongly as y2(4) and y3(8). Essentially for y2, 20 reports and for y3, 18 archives were effectively anticipated as y2 and y3 individually. From the Tables I and II the quantity of right expectations for every name was considered for assessing the framework. The quantity of right forecasts for marks y1,y2 and y3 for Method1 and Method 2 can be compressed as in Table III.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

***TABLE III : NUMBER OF CORRECTPREDICTIONS***

| Labels | Method 1 | Method 2 |
|---|---|---|
| y1 | 17 | 18 |
| y2 | 16 | 20 |
| y3 | 15 | 18 |

With this data we can speak to the outcome in a graphical arrangement which thinks about the quantity of right forecasts. The diagram looks at the forecasts of Method1 and Method2.

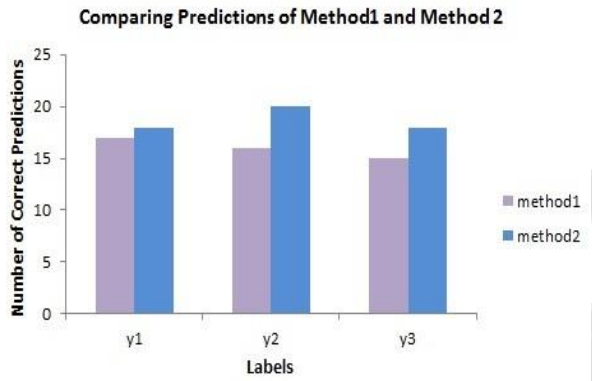


***Fig 2. Correct Predictions: Method 1 Vs Method 2***

As in Figure 2, the diagram demonstrates that that element choice approach with the current technique enhances the expectation accuracy in grouping.

•Analysis in view of Dimension and Time: Time unpredictability of a calculation evaluates the measure of time taken by a calculation to keep running as a component of the length of the string speaking to the data. Here in this work the time considered is the time taken for figuring or grouping of the testing input. The Method1(TESC) makes a term framework with every one of the terms in every one of the reports under each mark. In the wake of applying the

record recurrence thresholding i.e., Method2 (TESC+DF) the term lattice is made after an archive recurrence thresholding. By this technique the framework examinations every one of the terms and the quantity of reports in which the term shows up. An edge is given with the goal that the terms for which the archive recurrence is not as much as the edge esteem is expelled from the vector space. This causes the measurement diminishment of the vector space and accordingly the TF-IDF network. The fundamental favorable circumstances of this approach is

•Reduction of the vector space

•Identification of key highlights

•Less preparing time

Since the measurement of the vector space is lessened the time required for characterization is likewise decreased as it considers just the pertinent terms. So obviously the time and space many-sided quality of the proposed technique contrasted with the base strategy is decreased.

The general preparing and testing process is a tedious undertaking. Particularly with regards to the instance of content data the terms and classes might be corresponded so the characterization turns into a troublesome undertaking.

## 5. CONCLUSION

In this work the essential concentrate is on a successful content order system which embraces a semi-administered bunching approach. To make the base technique more compelling an element determination strategy was incorporated. The trials led on Reuter-21578 show preferable execution over the current technique.

The proposed framework would thus be able to be utilized as a part of different application spaces of content preparing. Multi-mark ideas are presently regularly utilized as a part of compelling content characterization procedures which are demonstrated strategy than thinking about a solitary name arrangement. This work can be enhanced by including ideas like outfit demonstrate for multi-name [16] order and more effective techniques for highlight determination which can be incorporated into the future examination. This proposition is beginning advance towards the objective of building a productive order display that can be utilized for tremendous data and in a multi-mark idea. There are degree for improvement and issues to be investigated for future examinations. The framework can be enhanced by fusing new ideas and techniques. Term recurrence

report recurrence or different upgraded strategies for highlight determination, Ensemble demonstrate, Voting system and so on can be fused for enhanced renditions.

## REFERENCES

[1]. Jiliang Tang and Salem Alelyani and Huan Liu, "Feature Selection for Classification: A Review",url: http://citeseerx.ist.psu.edu

[2]. Feng Chen, Pan Deng, JiafuWan and Daqiang Zhang and Athanasios V Vasilakos and Xiaohui Rong, "Data Mining for the Internet of Things:Literature Review and Challenges" , In Proc. of Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, 2015, url: http:// dx.doi .org/10. 1155/2015 /431047.

[3]. Nimit Kumar and Krishna Kummamuru, "Semisupervised Clustering with Metric Learning Using Relative Comparisons", IEEE Transactions on Knowledge and Data Engineering, vol,20, issue.4 April 2008.

[4]. Vikram Singh and Balwinder Saini, "An Effective Pre- Processing Algorithm for Data Retrieval Systems", International Journal of Database Management Systems - IJDMS, December 2014, vol.6,issue 6.

[5]. Wen Zhang, Taketoshi Yoshida and Xijin Tang, " A comparative study of TF-IDF LSI and multi-words for text classification", In Proc. of Expert Systems with Applications 38, pp. 2758– 2765, 2011, url: www.sciencedirect.com.

[6]. Yung-Shen Lin,Jung-Yi Jiang and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, vol.26 , pp. 1575–1590, July 2014.

[7]. Xiaofei Zhoua and Yue Hua, Li Guoa, "Text Categorization Based on Clustering Feature Selection", In Proc. of Procedia Computer Science 31, pp. 398-405, 2014, url: www.sciencedirect.com.

[8]. Yimming Yang and Jan O Pedersen,"A Comparitive study on Feature Selection in Text Categorization", 2012.

[9]. Yan Xu, Bin Wang, JinTao Li and Hongfang Jing, "An Extended Document Frequency Metric for FeatureSelection in Text Categorization", In Proc. of Springer- Ve rlag Berlin Heidelberg, 2008, pp.71-82.

[10]. Diederik P Kingma, Danilo J Rezendey, Shakir Mohamedy and Max Welling, "Semi-supervised Learning with Deep Generative Models", Proceedings of the International Conference on Machine Learning ICML,October 2014

[11]. Ishtiaq Ahmed, Rahman Ali, Donghai Guan, Young-Koo Lee, Sungyoung Lee and TaeChoong Chung, "Semi- supervised learning using frequent itemset and ensemble learning for SMS classification", In Proc. of Expert Systems with Applications, 2014.

[12]. Zhaocai Sun,Yun ming Ye, Xiaofeng Zhang, Zhexue Huang, Shudong Chen and Zhi Liu "Batch-Mode Active Learning With Se mi-supervised Cluster Tree For Text Classification", In Proc. of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012, pp.388-395.

[13]. Reuter-21578 Distribution 1.0 , Available Online at url: http://www.research.att.com/lewis.

[14]. Stanford POS Tagger, Available Online at url: http://nlp.stanford.edu/software/tagger.shtmlDownload.

[15]. TF-IDF calculation, Available Online at url: http://www.tfidf.com/

[16]. Cosine Similarity, Available Online at url : https://en.wikipedia.org/wiki/ Cosinesimilarity"

[17]. Wei Bi, James T Kwok, "Efficient Multi-label Classification with Many Labels", In Proc. of the 30 th International Conference on Machine Learning, vol. 28, 2013