"AUTOMATED TEXT DOCUMENT CLASSIFICATION USING PREDICTIVE NETWORK"

[1]Mr. Suraj S. Bhoite **[2]Ms. Swapnali K. Londhe**
Assistant Professor at the Department of Computer Engineering,
PES Modern College of Engineering, Pune-05

Abstract– *The automatic text classification methods are supposed to retrieve the RI as well as organize it according to it is the degree of relevancy with the given query. The main problem in organizing is to classify which document is relevant & which is relevant. The Automated text classification consists of automatically organizing clustered data. We propose an automatic method of text classification using a Machine, Automatic text classification using TFIDF, Word sense word embedding algorithm, and K-means clustering classification machine learning algorithm, firstly apply the pre-processing method and remove not needed data then evaluate TF-IDF and Semantic relations of every word and classify data using machine learning algorithm on student conference papers dataset. we use the word net word embedding algorithm to eliminate the ambiguity of words so that each word is replaced by its meaning in context. The closest ancestors of the senses of all the undamaged words in a given document are selected as classes for the specified document.*

Keywords- Machine Learning, Descriptive Clustering, Pre-processing, K-Means Clustering.

1. INTRODUCTION

The mass of data obtained from our increases. This data would be irrelevant if our capacity to productively get to didn't increment too. For most extreme advantage, there is a need for devices that permit look, sort, list, store and investigate the accessible information. One of the hopeful regions is automatic text classification. Envision ourselves within the sight of an impressive number of texts, which are all the more effectively available on the off chance that they are composed into classes as per their topic. One could request that humans read the text and arrange them physically. This assignment is hard if done on hundreds, even a huge number of texts. Thus, it appears to be important to have a computerized application, so here automatic text categorization is presented. Growing the number of Data Mining presentations include the complex & structured type of Data & Require the use of sensitive language.

Unfortunately Existing approach, especially by using Logic Programming methods, often suffer is not low scalability. When allocating with difficult Database schemas but also insufficient predictive performance usage noisy value in real-life applications. Though, levelling approaches is wont to need significant time & effort for the data conversion, outcome in behind the compressed representations of the standardized Database, & produce are extremely large data with huge Number of extra attribute value & frequent NULL values. In this system, the result is problems have prevented a diverse application of multi Relational Mining and challenges the Data Mining community. This article introduces a Descriptive clustering methodology where flattening is required to bridge the gap between propositional learning algorithms and relational.

In the proposed methodology, Data Analysis Technique, clustering it can classify a subset of information example with mutual characteristics. The operator can explore the information by examining the user can search the data by using selected examples in the group instead of relatively examining the instances of the complete data set. This allows users to focus efficiently on large relevant subsets of Data sets, in particular for document collections. In particular, the descriptive grouping consists of an automatic combination set of the parallel instance in the cluster and repeatedly generates for the explanation or synthesis that can be interpreted by man for each group. The description of each cluster allows a user to determine the relevance of the grouping to observe its content in the text documents. Quality is the grouping is important, so that it is aligned with the idea of the likeness of the user, but it is similarly imperative to deliver a user with a brief & useful instant that correctly replicates the constant of the cluster.

1.1 Motivation

In the initial part for the study of Text Classification, their application, which algorithm is used for that, and the special type of model, I resolute to work on the text classification which is used for data analysis work done on that purpose, and that the method used for that application is text classification using usual data mining algorithms. The analysis depends on three main objectives.

1. To giving labels to academic conference papers by technical domains and sub-domains.
2. To creating suggestions and Predictions using K-means clustering.
3. This system is used for automatic text classification.

2. RELATED WORK

A literature survey is the utmost significant step in any kind of investigation. Already start developing we need to study the previous papers of our domain which we are working on and based on of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly analyse the related work on automated text document classification, & their different techniques.

Austin J. Brockmeier, Member, IEEE, Tingting Mu, Member, IEEE, Sophia Ananiadou, and John Y. Goulermas, Senior Member, IEEE, "Self-Tuned Descriptive Document Clustering using a Predictive Network," 2018.

In this paper, cluster predictions are made using logistic regression models, and feature predictions rely on logistic or multinomial regression models. Optimizing these models leads to a completely self-tuned descriptive clustering approach that automatically selects the number of clusters and the number of features for each cluster. They applied the methodology to a variety of short text documents and showed that the selected clustering, as evidenced by the selected feature subsets, is associated with a meaningful topical organization.

K.Meena, R.Lawrance "An Automatic Text Document Classification using Modified Weight and Semantic Method," IJITEE Oct 2019.

This proposed method is used for descriptive type answer evaluation. Manual evaluation of descriptive-type paper may lead to a discrepancy in the mark. It is eliminated by using this type of evaluation. This system has been tested with answers written by learners of our department. It allows more accurate assessment and more effective evaluation of the learning process. This method has a lot of benefits such as reduced time and effort, efficient use of resources, reduced burden on the faculty, and increased reliability of results. The classification algorithm Support Vector Machine (SVM) is used to classify the dataset. The experimental analysis and results show that the performances of the proposed feature extraction methods are outperforming the existing feature extraction methods.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes and Donald Brown, "Text Classification Algorithms: A Survey," March 2019.

In this paper, a brief overview of text classification algorithms is discussed. This overview covers different text feature extractions, dimensionality reduction methods, existing algorithms and techniques, and evaluation methods. Finally, the limitations of each technique and their application in real-world problems are discussed.

J – T Chien, "Hierarchical Theme & Topic Modeling," IEEE Trans. 2016.

This paper is taking into account hierarchical data sets in the body of text, such as words, phrases, and documents, they perform Structural Learning and they deduce hidden themes, and themes for sentences and words from a group of documents, correspondingly. The association of the arguments &arguments in dissimilar Data groups are discovered through an unverified procedure without preventive the number of the clusters. They build a Hierarchical theme and a thematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and the thematic words are

extracted. The proposed method is evaluated as active for the construction of a Semantic Tree Structure of the corresponding Sentences and Words.

Bemardini, Carpineto, & M.D Amico, "Full- Subtopic with Key phrase – Based Searching Results Clustering," IEEE 2009.
Study of this problem of restoring many documents that are relevant to individual sub-topics gives the Web query, called "full child retrieval". They are using a new algorithm for grouping search results that generate clusters labelled with key phrases. The key phrases are removed general suffix tree created by the results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents". They have used a test collection specifically designed to evaluate the recovery of the sub-themes, they have found that our algorithm has passed both other clustering algorithms of present research outcomes as a method of redirecting search results underline the diversity of results.

C. Zhai& Q. Mei, "Automatic Libelling of Multinational Topic Models," Pro. AcmSigkddInt. Conf 2007.
In this paper, they suggest probabilistic methods to repeatedly labelling multinomial topic copies in our objective way. They are cast of labelling problem is an optimization problem involving reducing word distributions and exploiting mutual info between a label and a topic model. Experiments with user studies have been done on two text data sets with different genres. These system results are the proposed labelling methods are quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. Their methods are universal and can be applied to labelling topics learned through the kinds of topic models such as PLSA, LDA, and their variations.

R. Lotlikar, K. Kummamuru, K. Singal, R. Krishnapuram, "A Hierarchical Monothetic Document Clustering Algorithm for Summarization & Browsing Search Result, "Proc. Int. Conf 2004.
This paper Organizing Web search results in a hierarchy of themes and secondary themes make it easy to explore the collection and position the results of awareness. They are proposing a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to a query. At all levels of the hierarchy, the new algorithm increasingly finds problems to maximize coverage and maintain the distinctiveness of the topics. They discuss the algorithm proposed as Discover. The evaluation of the quality of a hierarchy of subjects is not a trivial task; the last test is the user's judgment.

D. Wunsch& R. xu, "Survey of Clustering Algorithms," IEEE Trans. 2005.
In this system, Data Analysis shows of crucial role inconsiderate several occurrences. Conglomerate Analysis, a primitive examination with little or no earlier knowledge, contains the research developed in a Wide Variety of communities. Diversity, on the one hand, provides us with many tools. On the other hand, the profusion of options confuses. They have examined the grouping Algorithm for the Data-Sets that appear in indicators, Computer Science & Machine Learning and they illuminate their uses in some reference Data-Set, the problematic of street vendors & Bioinformatics, and one area that attracts intense exertions.

T. Kohonen, S. Kaski, J. Honkela, A. Saarela, K. Lagus, "Self Organization of a Massive Document," IEEE Trans. 2000.
This paper defines the execution of the system that can organize large collections of documents based on written parallels. It's based upon the self-organized map (SOM) algorithm. Like the feature courses for documents, the arithmetical symbols of their words are used. The important objective in this work was to resize the SOM Algorithm to handle large amounts of high-dimensional data.

3. PROPOSED APPROACHES

In this system, we propose automatic text classification using TFIDF, Word sense word embedding algorithm, and K-means clustering classification machine learning algorithm, firstly apply the pre-processing

method and remove not needed data then evaluate TF-IDF and Semantic relations of every word and classify data using machine learning algorithm on student conference papers dataset.

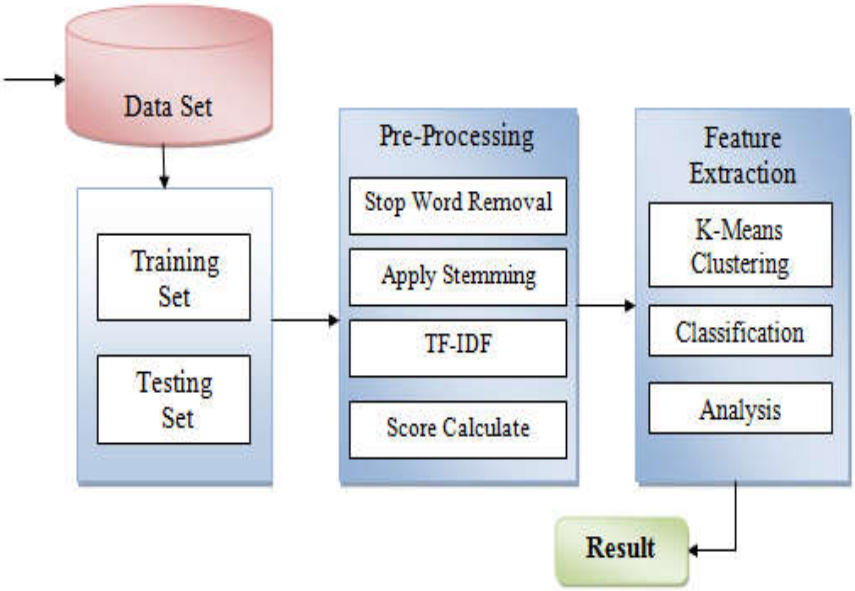


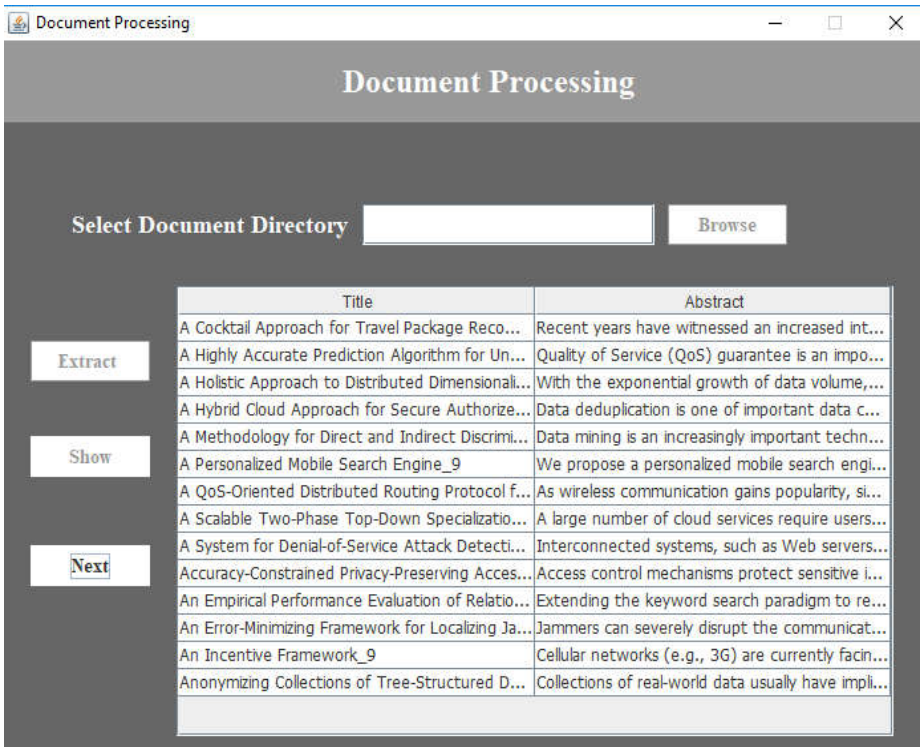
Fig1. System Architecture

Step 1. Dataset-

The Dataset is divided into two modules one is the training module and another is the testing module.

1) Training Module:-

The Training Module firstly browses the conference paper dataset for train dataset using pre-processing algorithm including stop word removal, tokenization, and stemming. After that calculate the score of a dataset by using the TF-IDF Algorithm. Only an admin can train the dataset.

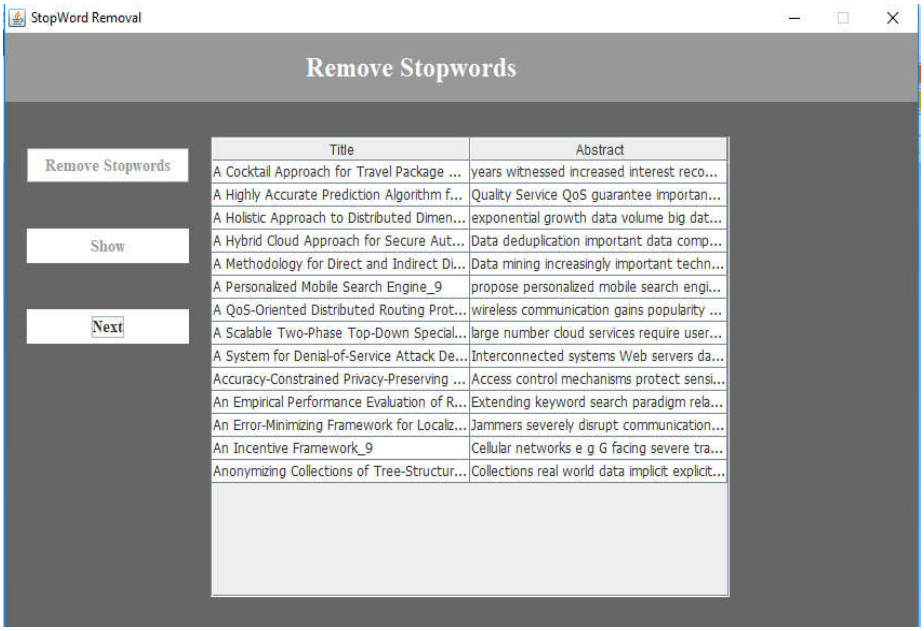


2) Testing Module:-
The testing Module will browse the dataset trained by the admin for finding the conference paper domain using the classification of the K-means Algorithm.

Step 2. Reprocessing-
Data Pre-processing is the data is processed beforehand to remove errors and improve efficiency and accuracy by using tokenization- the process of breaking a text stream into words, symbols, phrases. Stop Words Removal is removing the word like is, are, they, but, etc. stemming remove suffix and prefix.

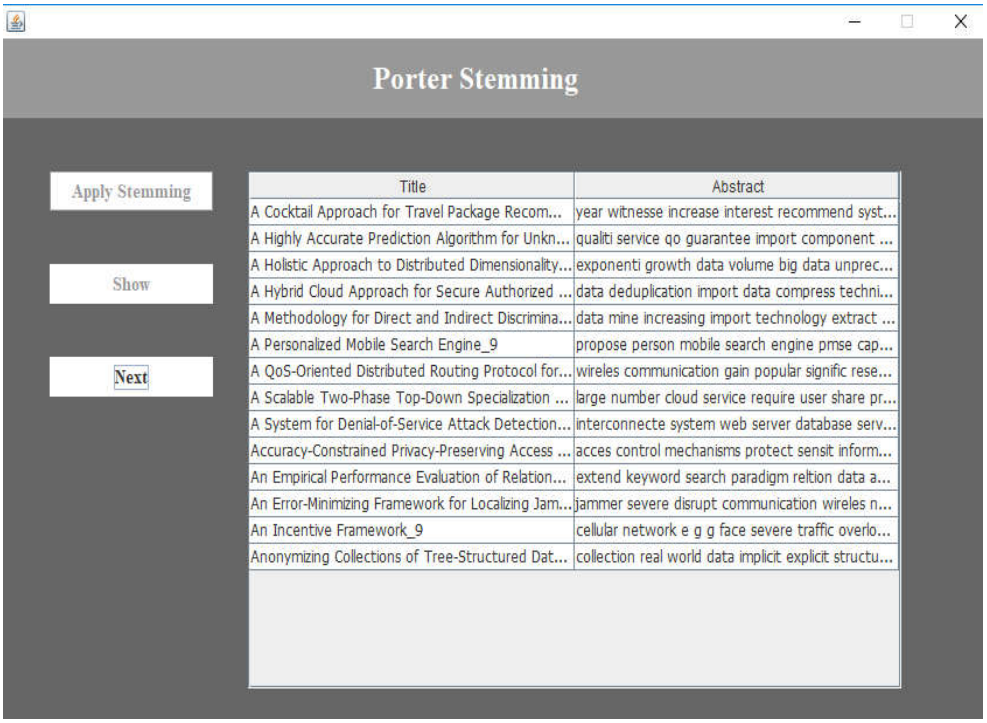
A] Tokenization: -
Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other significant elements called tokens. This technique removes Special characters and images. Like that; #, *, $, @ etc.

B] Stop Word Removal: -
A lot of words in documents return too repeated but are mostly worthless as they are used to join words together in a sentence. It is generally understood that stop words do not contribute to the context or content of textual documents. Stop words are very regularly used common words like *'and', 'are', 'this'* etc. They are not helpful in the classification of documents. So they must be removed. By making a list of stop words, we can exclude them in the subsequent text analysis.



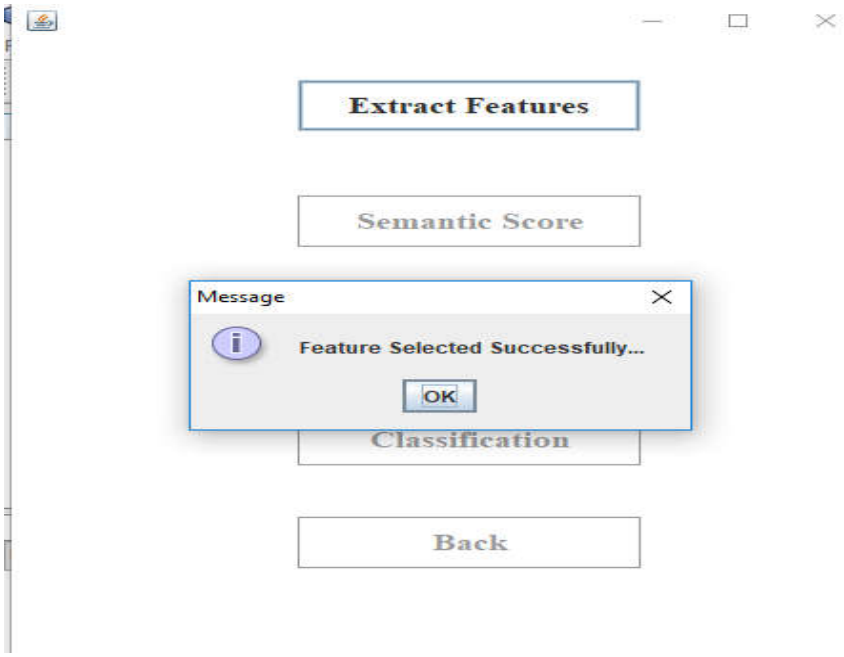
C] Stemming Remove: -
Stemming is the process of conflating the variant forms of a word into a frequently represent the stem. This algorithm process is semantic equivalence, in which the Different ways of the word the reduced to the mutual arrangements. This technique removes suffix and prefix and finds the original words. For E.g.;

Form	Suffix	Stem
Plays	-s	Play
Played	-ed	Play
Playing	-ing	Play
Studies	-es	Study



Step 2: Feature Extraction-

The feature extraction reduces the no. of features in a dataset by creating new features from the existing dataset. These newly reduced features will summarize much of the information contained in the original set of features. The common techniques of feature extractions are Term Frequency-Inverse Document Frequency (TF-IDF). After feature extraction calculates the semantic score of the dataset.



4. ALGORITHMS AND MATHEMATICAL MODEL

1. **K-Means Algorithm:**
K-Means is one of the unsupervised learning algorithms for clusters. Clustering the image is grouping the pixels according to the same characteristics. In the k-means algorithm firstly we have to define the number of clusters k. Then the k-cluster center is chosen randomly. The distance between each pixel to each cluster center is calculated. The distance may be of simple Euclidean function. A single-pixel is compared to all cluster centers using the distance formula. The pixel is moved to the particular cluster which has the shortest distance of all. Then the centroid is re-estimated. Again each pixel is compared to all centroids. The process continues until the center converges.

2. **TF-IDF Algorithm:**
The TF-IDF score is term I_{ij} is calculated by the Term Frequency & Inverse Document Frequency. The TF & IDF is the fundamentals of the most outstanding general term weighting system in IR.

TF-IDF = TF (Term Frequency) *** IDF** (Inverse Document Frequency)
TF:- TF measures the frequency of a term (t) in a document. It is given by-

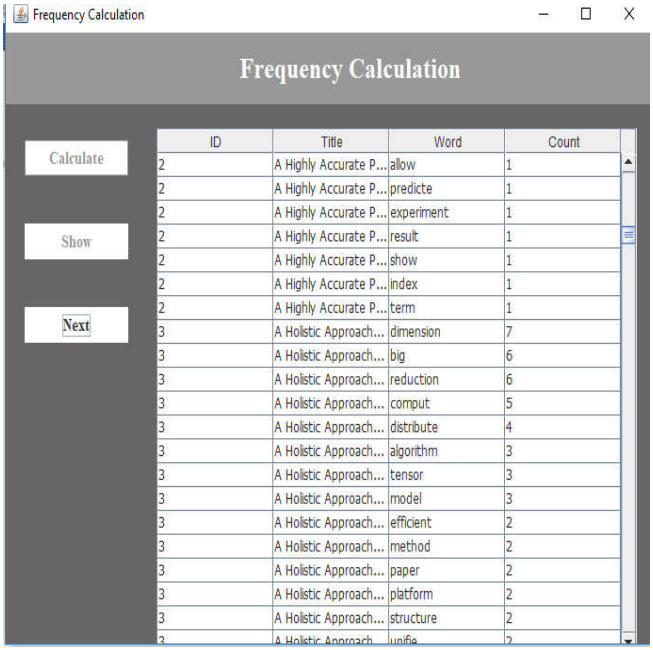
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

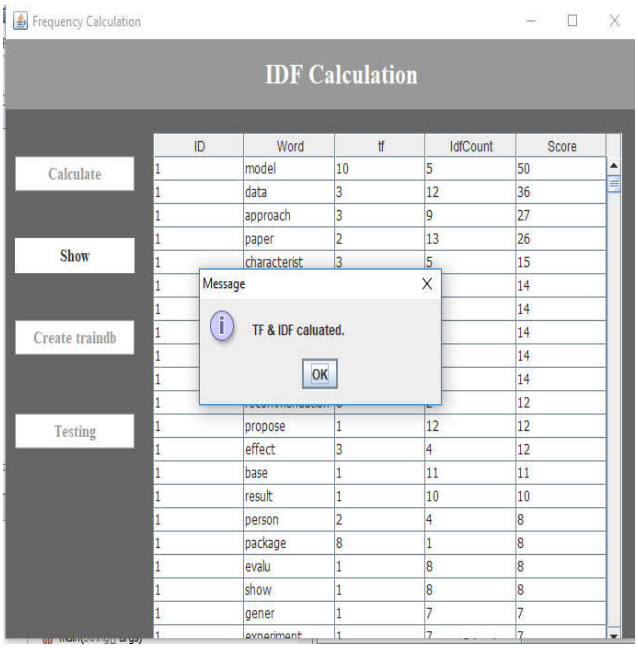
Where,
$\mathbf{n_{i,j}}$**:-** frequency of occurrence of t_i in document d_j
$\sum_k \mathbf{n_{k,j}}$**: -** the sum of the frequency of all words in d_j

IDF: - TF gives equal importance to all words but IDF Measures how important a Word is IDF can be computed using given below-

$$idf_i = \text{Log} \frac{|D|}{|\{d; d \in t_i\}|}$$

Where,
D: Total Number of Documents
d: Number of documents containing t_i.





➢ **Mathematical Model**

Let us consider S as a System for Document Classification.
S= *{I, F, O}*
Where
I = Text Dataset Uploaded by the User
F = Functions Implemented to get the Output
O = Output i.e. Document Clustering

Input:-
Classify the Inputs:
Set,
$\boldsymbol{F}= \{F_1, F_2, F_3 \ldots\ldots F_n\}$
Where
F is a Set of the Function to Execute Commands.
$\boldsymbol{I}= \{I_1, I2, I3 \ldots\ldots I_n\}$
Where *I* is set of input to the function set.
$\boldsymbol{O} = \{O_1, O2, O3 \ldots\ldots\ldots O_n\}$
Where *O* is set of output from the function set.

5. RESULT

This system result, we analyse this system is based on the student conference papers dataset this available on the internet. The model valuation metrics used in this system are accuracy, precision, recall, and F1 score. The calculation parameters are defined as follows:

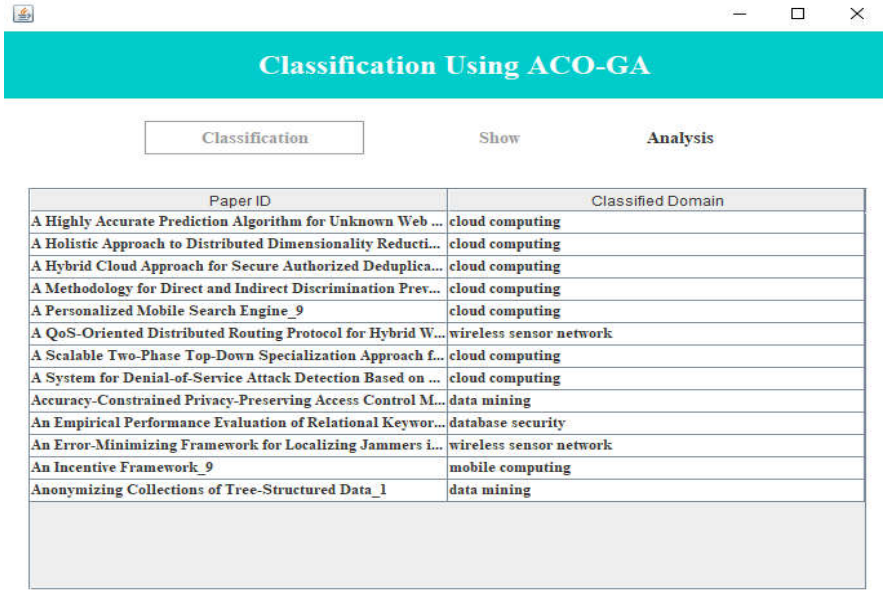
TP: True Positive (properly expected no. of instance)
FP: False Positive (wrongly expected no. of instance)
TN: True Negative (properly expected no. of instance as not required)
FN: False Negative (wrongly expected no. of instance as not required)

$$\boldsymbol{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$\boldsymbol{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \boldsymbol{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \boldsymbol{F1} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

		Actual	
		Positive	***Negative***
Predictive	***Positive***	True Positive	False Positive
	Negative	False Negative	True Negative

A. Analysis of Conference Papers Domain Detection

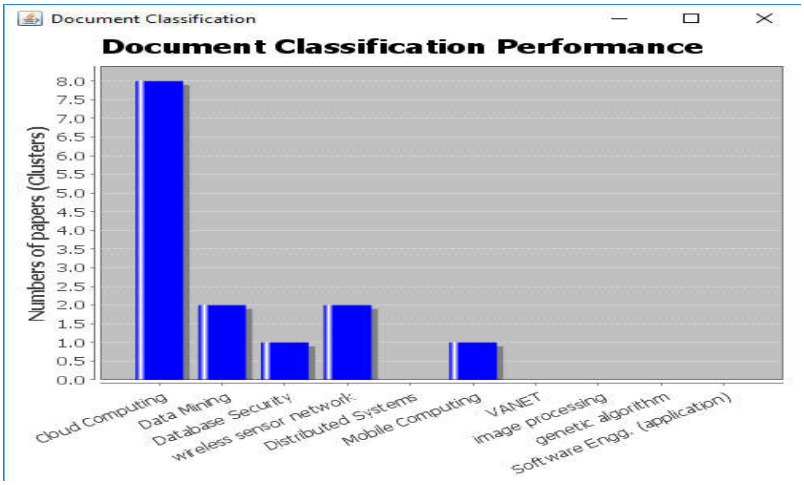


B. Domain Detection Result

Sr. No.	Domain Name	Paper Coun
1	Cloud Computing	07
2	Data Mining	02
3	Database Security	01
4	Wireless Sensor Network	02
5	Distributed System	00
6	Mobile Computing	01
7	VANET	00
8	Image Processing	00
9	Genetic Algorithm	00
10	Software engineering	00
	Total Papers	**13**

Table 2: Domain Detection Result

C. Accuracy Graph



6. CONCLUSION AND FUTURE WORK

In this system, we propose automatic text classification using TF-IDF, and K-Means Clustering Classification machine learning algorithm, firstly apply the pre-processing technique and remove unwanted data then calculate score using TF-IDF and semantic relations of every word and classify data using Machine Learning Algorithm on student conference papers dataset. Feature extraction methods, such as TF-IDF, are commonly used in both academic and commercial applications. For future work, plan to investigate our framework using more complex features including multi-word expressions, named entities, and clusters of features themselves.

REFERENCES

1. Austin J. Brockmeier, Member, IEEE, Tingting Mu, Member, IEEE, Sophia Ananiadou, and John Y. Goulermas, Senior Member, IEEE, "Self-Tuned Descriptive Document Clustering using a Predictive Network," 2018.

2. K.Meena, R.Lawrance "An Automatic Text Document Classification using Modified Weight and Semantic Method," IJITEE Oct 2019.

3. Kamran Kowsari , Kiana Jafari Meimandi , Mojtaba Heidarysafa , Sanjana Mendu , Laura Barnes and Donald Brown, "Text Classification Algorithms: A Survey," March 2019.

4. J – T Chien, "Hierarchical Theme & Topic Modeling," IEEE Trans. 2016.

5. Bemardini, Carpineto, & M.D Amico, "Full- Subtopic with Key phrase – Based Searching Results Clustering," IEEE 2009.

6. C. Zhai& Q. Mei, "Automatic Labeling of Multinational Topic Models," Pro. AcmSigkddInt .Conf 2007.

7. R. Lotlikar, K. Kummamuru, K. Singal, R. Krishnapuram, "A Hierarchical Monothetic Document Clustering Algorithm for Summarization & Browsing Search Result, "Proc. Int. Conf 2004.

8. D. Wunsch& R. xu, "Survey of Clustering Algorithms," IEEE Trans. 2005.

9. T. Kohonen, S. Kaski, J. Honkela, A. Saarela, K. Lagus, "Self Organization of a Massive Document," IEEE Trans. 2000.