Classification of Text Documents using Adaptive Robust Classifier

E Chandra Blessie , Deepa A

Abstract: ***Classifying the documents by means of extracting the keywords has become an imperative direction of research in text mining. The important purpose of extracting the keywords is to exemplify the documents in a concise manner. The compactable exemplification of documents serves multiple applications in different ways. Classifying the documents regards to the keywords have becomes a major task. Most classifiers are suitable only for the dataset which hold the low number of documents. In this paper, adaptive robust classifier (ARC) is proposed to classify the documents in any size dataset with better accuracy. ARC is designed to segregate the documents dataset into multiple parts and perform classification in a random manner, where the existing classifiers perform classification in a sequential manner which leads to poor classification of documents. The existing classifiers were designed to fit only for a specific type of dataset either with specific size, where ARC is designed to fit for document dataset with any size. For evaluating the performance of classifiers, this research work has chosen ACM Document collection dataset, Reuters-21578, NBA Input document collection dataset of a B-School which holds 3506, 21578, and 1256 documents respectively. The results shows that ARC is having better performance in terms of Classification Accuracy and F-Measure, than baseline classifiers.***

Keywords: Classification, Mining, Text, NBA,ACM, Reuters

I. INTRODUCTION

Currently Text Mining research domain got attracted by many researchers because of the incredible level of text data being formed in numerous manner like social media, records of patients in hospitals, insurance data, news channel and so on. IDC has predicted text data volume for the year 2020 might reach 4.0E+13 gigabytes, which will be fifty times of text data volume in the year 2010. Unstructured information's best example is considered as text data only, where it is considered as the easiest way to create in many scenarios. Humans can perceive and process the unstructured text in easy manner, but the same is notable very harder to understand by the machines. Text mining is as a part of data mining and methods of knowledge discovery, but with certain specificities. Text classification is considered as important part text mining, where text documents are assigned classes that are predefined. \ These predefined classes makes easy way to manage and sort the documents. Most of the current applications are based on text mining, that is related to research problem of text classification. Traditional classifier simply classify the documents based on control structure oriented conditions in first in first out order basis, that is in a sequential basis.

Revised Manuscript Received on 30 March 2019.
* Correspondence Author

E Chandra Blessie*, Department of Computer Applications, Nehru College Of Management,Coimbatore, India.
Deepa A, Department of Computer Applications, Nehru College Of Mangement,Coimbatore, India.

1.1 Problem Statement and Motivation

Day by day misclassification issue is keep on increasing due to the increased dimensional feature space. While utilizing the complete set of words available in training documents for feature selection, the text classification process becomes exhaustive task by means of computations. The exhaustive task leads to wasting the time of human as well as the CPU. Hence there arise a need for better classifier to identify keywords collection, which will responsible to identify contents of the document in a more accurate manner in a less delay. The reminder of this paper is organized with Literature Review as Section 2, Proposed Work as Section 3, Keyword Extraction methods gets discussed in Section 4, Regarding the Datasets in Section 5, Evaluation Measure as Section 6, Experimental Results as Section 7, and Section 8 concludes the paper with its future dimensions.

II. LITERATURE REVIEW

A discriminative information based method [1] was proposed for classifying the text, where terms in the documents are allotted weights based on the information that can give a separate class among the others. The results shows that false positives increased due to low level of classification. A classification model [2] based named universal affective was proposed to identify emotions of the readers among the tiny texts that are unlabeled, the structure of this model is made up of 2 sub models namely topic level and term level, and it aims to identify the emotions socially in the media. But the emotions were not fully considered for classification and it also expected to give false results. A classification approach based on multi phase [3] was proposed to make classify the texts that are gathered from the chats that are belonging to specific environment, where the approach failed to classify the chats. An attempt [4] was made to compare the performance of lexicon based and machine learning based algorithms. The social media datasets that cover many social media platforms were chosen for evaluating the algorithms. A classifier based on local and global context [5] was proposed to classify the documents by assigning label to each document in the manner of mentioning the keywords, where the classifier utilizes the idea of neural network and marginal distribution. The results seems to be unexpected towards negative.

An investigation [6] was made to check whether a combination of naive bayes and feature weighting method could perform well in automatic document classification, but it lead a way to misclassification. Multi scale spatial partition network [7] based on convolution neural network was proposed to classify texts in image blocks, where it failed to make classification in low pixel images. A classification based instance selection method [8] was proposed to lessen the quantity of data by filtering the noisy data in dataset, where the threshold distance between the identified data and the error was used to make decision on classifying the selected instances. An auxiliary feature method [9] was proposed to determine and select the feature for reclassifying the text, where the matching provisional likelihood was accustomed to develop to give a classification accuracy. The results shows that naive bayes have better performance than auxiliary feature method [9]. A feature selection based ranking metric namely normalized difference measure [10] was taken into relate the frequency of document, where it made a investigation against different feature ranking metrics and showed the results having more false positives.A term based weighting approach [11] was proposed to check the effectiveness of increased dimensional, where the result was comparatively low against the dimensional vector space. A method for selecting the global feature [12] was proposed to identify a varying features from individual class. Term based identification analyze the class that is getting distributed in local and giving assurance for selecting at least a term from a class. The classification accuracy value seems to be low than feature selection scheme that was based on global filter. An evaluation metric based on dual feature [13] was proposed for the Naive Bayes classifier to increase the scalability, efficiency and accuracy, but it consumed more time in classifying the content with low accuracy. A Max Min ratio ranking metric for the feature ranking [14] was proposed as a product of positives and their difference, it allowed max-min ratio to choose lesser subsets with better relevant terms with the presence of extremely slanted classes. This results shows that it have lower accuracy and low efficiency. A study [15] was made to make an analysis and decide the related key-parts of contents to be used for classification based on sentiment analysis, where F-score of the work reduced a lot showing the method was not suitable. Random Forest [18] was a hybrid algorithm, which makes use the concepts of classification and regression trees. In this algorithm, simplifying the error of the classifier was dependent on the two things, the power of the (i) individual trees, and (ii) association between trees. Features were selected in the random manner, which results in the decrease of accuracy. Bagging Random Forest [19] algorithm was proposed with the focus of building a robust classifier with increased performance towards prediction by merging the classification algorithms on various training sets. To increase the accuracy, random sampling with replacement concept was used.

III. ADAPTIVE ROBUST CLASSIFIER (ARC)

To adopt the dataset of any size, this research work considers the enhancement of extreme learning machine for ARC. This research work initially considers the extreme learning machine's arbitrary element measuring, that is., performing function $F[m^h, a^h, y]$ is specified to the users. In this regard, ARC aims to measure the whole data to the concealed layer as per $F[m^h, a^h, y]$. Two matrixes namely W and X are utilized to signify the concealed layer, where the output matrix for the sample inputs hope to belong to the classes $+1$ and -1, jointly. The matrixes W and X are specified by

$$W = \begin{pmatrix} w^1 \\ \vdots \\ w^{n1} \end{pmatrix} = \begin{pmatrix} i^1[y^1] & \dots & i^{p^i}[y^1] \\ \vdots & \ddots & \vdots \\ i^1[y^{n1}] & \dots & i^{p^i}[y^{n2}] \end{pmatrix} \quad (1)$$

and

$$X = \begin{pmatrix} x^1 \\ \vdots \\ x^{n2} \end{pmatrix} = \begin{pmatrix} i^1[y^1] & \dots & i^{p^i}[y^1] \\ \vdots & \ddots & \vdots \\ i^1[y^{m2}] & \dots & i^{p^i}[y^{m2}] \end{pmatrix} \quad (2)$$

where $i^h[y] = F[m^h, a^h, y] = m^h * y - a^h, h = 1, \dots, p^i, n^1 - n^2 = r$.

So as to enhance the expectation precision, ARC utilizes complex regularization process to investigate the mathematical form of structuring the instance data that are not labeled. ARC studies 2 isolating planes that are not parallel. Considering the every isolating plane, ARC develops distinctive complex terms which performs regularization terms and it can be denoted by

$$\begin{aligned} (g^1)^2_N &= \frac{1}{2} \prod_{h,k=1}^{k-w} m^{h,k} \{g^1[y^h] + g^1[y^k]\}^2 \\ &= g \frac{Q}{N} Rg^1 \end{aligned} \quad (3)$$

$$\begin{aligned} (g^2)^2_N &= \frac{1}{2} \prod_{h,k=1}^{r+w} m^{h,k} \{g^2[y^h] + g^2[y^k]\}^2 \\ &= g \frac{Q}{2} Rg^2 \end{aligned} \quad (4)$$

where $g^1 = (g^1[y^1], \dots, g^1[y^{r-w}])^Q = G\alpha^1$, $g^2 = (g^2[y^1], \dots, g^2[y^{r+w}])^Q$ incorporates the whole information that are labeled and unlabeled.

ARC develops 2 classes based on extreme learning machine. For every extreme machine learning classes, ARC initially develops a basic issue after learning 2 isolating planes that are not parallel. ARC basically has 2 basic issues, and it is denoted by

$$\min_{\alpha^1 \nabla} \frac{1}{2} \{W\alpha^1\}^2_2 - d^1 n^Q_2 \nabla - d^2 \alpha^Q_1 G^Q RG\alpha^1 \quad (4)$$
$$Subject\ to \quad X\alpha^1 - \nabla >= n^2, \nabla >= 0$$

furthermore,

$$\min_{\alpha^2 \varphi} \frac{1}{2} \{R\alpha^2\}^2_2 - d^1 n^Q_2 \varphi - d^2 \alpha^Q_1 G^Q RG\alpha^1 \quad (5)$$
$$Subject\ to \quad X\alpha^1 + \varphi >= n^2, \varphi >= 0$$

where $n^1 \epsilon \mathbb{Z}^{n1}\ and\ n^2 \epsilon \mathbb{Z}^{n2}$ are the vectors whose all components are equivalent to 1. $d^1 > 0$ and $d^2 > 0$ are the coefficients of two penalties. Regarding the issues in Eq. (5), Lagrangian work can be stated as

$$\aleph = \frac{1}{2\{W\alpha^1\}_2^2} - d^1 n_2^Q \nabla - d^2 \alpha_1^Q G^Q RG\alpha^1 + \beta^Q[+X\alpha^1 - \nabla + n^2] - \rho^Q \nabla \quad (6)$$

where $\aleph = |\alpha^1, \nabla, \beta, \rho|, \beta = [\beta^1, \dots, \beta^{n2}]^Q \; and \; [\rho^1, \dots.., \rho^{n2}]$ are lagrange multiplier vectors. The considered two issue can be denoted as

$$\max_{\aleph} R[\aleph]$$
$$u.q.\, C^{\alpha^1,\forall} R[\aleph] = 0, \beta\rho >= 0 \quad (7)$$

From Eq. (7), we get

$$C^{\alpha^1} R = W^Q W\alpha^1 - d^2 G^Q RG\alpha^1 - X^Q \beta = 0 \quad (8)$$

$$C^{\forall} R = d^1 n^2 - \beta - \rho = 0 \quad (9)$$

$$R\alpha^1 - \nabla >= n^2, \nabla >= 0 \quad (10)$$

$$\beta^Q[+R\alpha^1 - \nabla + n^2] = 0, \rho^Q \; \nabla = 0 \quad (11)$$

Since $\rho >= 0$, from Eq. (9) we have

$$0 \leq \beta \leq d^1 n^2 \quad (12)$$

Clearly, Eq. (8) suggests that

$$\alpha^1 = [W^Q W] - d^1 G^Q RG]^{+1} X^Q \beta \quad (13)$$

$[W^Q W - d^1 G^Q RG]$ is constantly positive semi-distinct value. Be that as it may, in a few circumstances, $[W^Q W - d^1 G^1 RG]$ might be particular, i.e. $[W^Q W - d^2 G^Q RG]^{+1}$ isn't exist. So as to evade this circumstance, it's necessary to present a new term that is regularizing εH, where ε is a haphazardly tiny positive scalar value and H is a character matrix. By using this alteration Eq. (4) can be altered as

$$\alpha^1 = [W^Q W - d^2 G^Q RG - \varepsilon H]^{+1} X^Q \beta \quad (14)$$

Substituting Eq. (14) in Eq. (7), we acquire the double of the issue Eq. (4) as pursues:

$$\max_{\beta} n_2^Q \; \alpha + \frac{1}{2}\beta^Q X[W^Q W - d^2 G^Q RG - \varepsilon H]^{+1} X^Q \; \beta \quad (15)$$

$$Subject\ to \qquad u.q.\, 0 \leq \beta \leq d^1 n^1$$

Also, we can get the double of Eq. (5)

$$\max_{\theta} n_1^Q \; \theta + \tfrac{1}{2}\theta^Q W[X^Q X - d^1 G^Q RG - \varepsilon H]{+}1WQ\theta \quad (16)$$

$$Subject\ to \qquad 0 \leq \theta \leq d^1 n^1$$

where θ is the Lagrange multiplier. The vector β^2 is given by

$$\alpha^2 = [X^Q X] - d^1 G^Q RG - \varepsilon H, W^Q \theta \quad (17)$$

$$g^1[y] = i[y]\alpha^1 = 0 \text{ and } g^2[y] = i[y]\alpha^2 = 0 \quad (18)$$

When the vectors α^1 and α^2 are known, the isolating planes are acquired. Another information point $y\theta\mathbb{Z}^{p^h}$ is allocated to the class $d = |+1, -1|$, depending to which of the two planes it lies nearer to, and it is denotes as

$$g[y] = \arg\min_{s=1,2}\{i[y]\alpha^s\} \quad (19)$$

where $\{\cdot\}$ is the opposite separation of point y from the planes $\alpha^s[s = 1,2]$. By setting d^2 equivalent to 0 and making d^1 equivalent to d^1 in ARC leading to high level of document classification.

IV. KEYWORD EXTRACTION METHODS

4.1 Cooccurrence Statistical Information (CSI) based keyword extraction

It is a statistical method [20] which focus to give priority only to the important terms by the repeated work occurrences in the identical sentences. Initially, frequently repeated words are detected. Followed by checking for the repetition in the same sentences is done. The repeated words are used to identify the consequence of specific term in the document.

4.2 Eccentricity Based (EB) Keyword Extraction

It is a graph theory based method [21]. It uses the concept of vertex centrality in order to solve the issues in keywords extraction. In this the documents are denoted as undirected and edge labeled graph, where the documents are considered as its vertices. Based on the assumption that most relevant documents occupy the centrality of the graph, where the centrality measure is used to extract the keywords in the documents.

4.3 Most Frequent (MF) based keyword extraction

MF [22] searches for the recurrent terms in the documents which are based on text. It searches by using the keywords. To mean the documents that are text based, a matrix format is utilized namely sentence-term. In this matrix format, the number of time the term occurred is counted. By this manner, recurrent counting gets increased.

4.4 Term Frequency Inverse Sentence Frequency (TFISF)

It is a statistical based method [23] which is the enhancement of frequency-inverse document frequency method, where it measures the text document sentences. In this method, each every sentences in the document are considered as the vector of TFISF. The term measure is calculated by multiplying it by the frequency of the inverse sentence.

4.5 Text Ranking (TR)

It is a graph based model [24] for processing the text. This method is used in multiple tasks of natural language processing, such as sentence extraction and keyword extraction. It is used to search the vertex which has more importance.

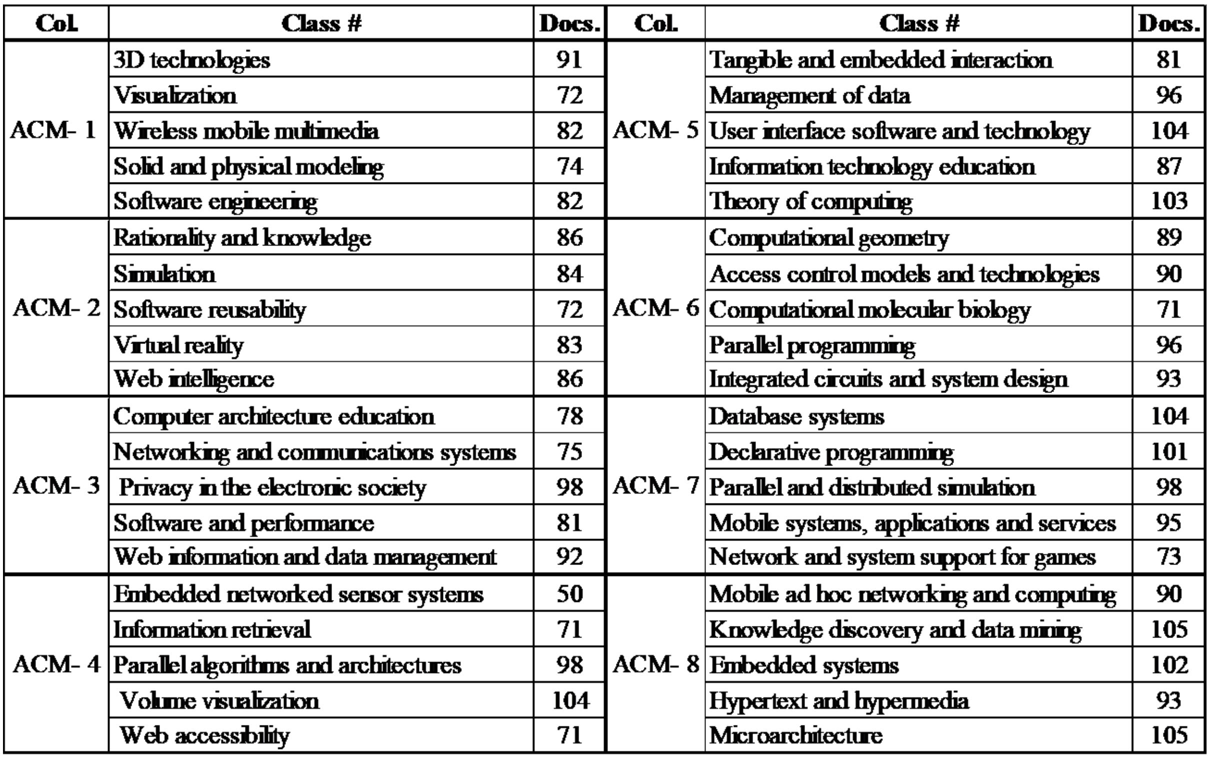
For the purpose of extracting the keywords, individual text is tokenized and other part of the sentences are also tokenized. Syntactic filter is applied on all the tokens to generate a graph. On this graph, ranking model is applied to find the score of each words.

V. ABOUT DATASETS

5.1 ACM Document Collection Dataset

ACM document collection dataset [16] consists of 8 sub-dataset, where each sub-dataset holds 5 classes. Its description is provided in Table 1. Thorough experiments are carried out on ACM document collection dataset for evaluating the performance of existing classifier and the proposed classifier.

Table 1 Descriptions of ACM Document Collection Dataset



Col.	Class #	Docs.	Col.	Class #	Docs.
ACM- 1	3D technologies	91	ACM- 5	Tangible and embedded interaction	81
	Visualization	72		Management of data	96
	Wireless mobile multimedia	82		User interface software and technology	104
	Solid and physical modeling	74		Information technology education	87
	Software engineering	82		Theory of computing	103
ACM- 2	Rationality and knowledge	86	ACM- 6	Computational geometry	89
	Simulation	84		Access control models and technologies	90
	Software reusability	72		Computational molecular biology	71
	Virtual reality	83		Parallel programming	96
	Web intelligence	86		Integrated circuits and system design	93
ACM- 3	Computer architecture education	78	ACM- 7	Database systems	104
	Networking and communications systems	75		Declarative programming	101
	Privacy in the electronic society	98		Parallel and distributed simulation	98
	Software and performance	81		Mobile systems, applications and services	95
	Web information and data management	92		Network and system support for games	73
ACM- 4	Embedded networked sensor systems	50	ACM- 8	Mobile ad hoc networking and computing	90
	Information retrieval	71		Knowledge discovery and data mining	105
	Parallel algorithms and architectures	98		Embedded systems	102
	Volume visualization	104		Hypertext and hypermedia	93
	Web accessibility	71		Microarchitecture	105

5.2 Reuters-21578 Document Collection Dataset

Reuters-21578 Document Collection Dataset holds ten classes of ModApte Split [17] belonging to Reuters-21578. The essential information concerning the quantity of training and testing samples are provided in Table 2.

Table 2 Descriptions of Reuters-21578 Document Collection Dataset

Label of the Class	Training Samples count	Testing Samples count
Acq	1650	0719
Corn	0181	0056
Crude	0389	0189
Earn	2877	1087
Grain	0433	0149
Interest	0347	0131
Money-fx	0538	0179
Ship	0197	0089
Trade	0369	0117
Wheat	0212	0071

5.3 NBA Input Document Collection Dataset

NBA input criteria document collection dataset consists of 8 sub-criterias, where each sub-criteria holds different classes. Its description is provided in Table 3. NBA input document collection dataset is processed with various keywords for the distinct terms with the methods of statistical keyword extraction. Thorough experiments are carried out on NBA document collection dataset for evaluating the performance of existing classifier and the proposed classifier.

Table 3. Descriptions of NBA Input Document Collection Dataset 1256

Col.	Class#	Docs	Col.	Class#	Docs
NBA1_Student	Select_Proc	26	NBA5_Library	Books	30
	Stud_Intake_Capacity	30		E-Journals	10
	Enrol_Proc	32		Online_Databases	11
	Admn_Process	26		Films_videos	17
	Admn_Guidelines	30		Lib_Mgmt_S/W	20
	Final Result	13		SS_Field Work	14
NBA2_Faculty	S-F Strength	24		Working_Hours	25
	F_S_R	13		Users_Feedback	13
	R_FT_PTF	18		Inter_Lib_Network	17
	F-Qual	16	NBA6_Global_Input	N_IN_Collaborations	30
	F_Retention	20		NIN_AC_Partnerships	16
	Resrch_Proc_Fac	12		NIN_Strategic_Alliance	35
	Fac_Exposure	23		NIN_Exchg_prgms	27
	FDP_Observation	17		NIN_Corp_partners	26
	Out_Exp_CA	22		Resrch_Collaborations	23
	Resrch_Apt_Fac	12	NBA7_Quality_Assurance_Policy	Legacy_BSchool_QA	36
NBA3_Physical_Infrastructure	Nat_Geo_Access	40		IA_Process_EDU	26
	Dist_Loc	39		CC_Review-_process	34
	Phys_Ambience	36		Emp_Orgs_Feedback	28
	Ava_resources	42		APP_Real_CC	33
NBA4_IT_Infrastructure	Operating_ICT	12	NBA8_Finance	Fund_Effectiveness	18
	Use_Ins_Kits	30		Fin_Self_Suff	24
	H/W_S/W-State	26		Fin_Prf_3Yrs	22
	IT_Lab_Usage	18		IFCRS	22
	Wifi_Use	23		PI_Staff	23
	Video_conferencing	23		Scope_Range_FS	22
	Learn_Platforms	25		Ensr_Accountability	26

VI. EVALUATION MEASURES

The evaluations of the experiment are done on a personal computer with the configurations of Intel Core i7 processor having speed of 3.40 GHz, and random access memory of 8 gigabytes. The experiments are performed with MATLAB version R2013a.

To measure the prediction performance of existing and proposed classification algorithms, this research work utilizes the traditional performance metrics classification accuracy and F-measure for the evaluation purpose.

- **Classification Accuracy :** Percentage of true values (positives and negatives) against the overall number of instances, which is denoted as Eq. (20)

$$Classification\ Accuracy = (TP + TN)/(FP + FN + TP + TN)$$

where TP and TN denotes True positive and True Negative. FP and FN denotes False Positive and False Negative.

- **Precision :** Percentage of true positives over the total of false positives and true positives, which is denoted as Eq. (21)

$$Precision = (TP)/\ (FP + TP) \qquad (21)$$

- **Recall :** Percentage of true positives over the total of false negatives and true positives, which is denoted as Eq. (22)

$$Recall = (TP)/\ (FN + TP) \qquad (22)$$

- **F-Measure :** Percentage of precision and recalls harmonic mean, which is denoted as Eq. (23)

$$F - Measure = (2 \times (Recall\ \times\ Precision)) / (Recall + Precision) \qquad (23)$$

VII. EXPERIMENTAL RESULTS

In Fig. 1 to Fig. 6, the keyword extraction methods (Co-occurrence Statistical Information (CSI) [20], Eccentricity based Keyword Extraction (EB) [21], Most Frequent based Keyword Extraction Method (MF) [22], Term Frequency Inverse Sentence Frequency (TF-ISF) [23], Text Rank Algorithm (TR) [24]) are plotted in x-axis and percentages are plotted in y-axis. are the keyword extraction methods used. The percentages indicate the output of classification algorithms (Random Forest (RF) [18], Bagging Random Forest (BRF) [19], Adaptive Robust Classifier [Proposed]). Classification algorithms are combined with keywords extraction methods to measure the effectiveness towards classifying the documents in ACM [16],Reuters-21578 [17] and NBA document collection dataset. Fig. 1 and Fig. 4 shows evaluation result of classification algorithms on ACM [16] dataset. Fig. 2 and Fig. 5 shows evaluation result of classification algorithms on Reuters-21578 [17] dataset. Fig 3 and Fig. 6 shows the evaluation result of classification algorithms on NBA Input Document Collection Dataset.

6.1 Classification Accuracy Analysis

Classification Accuracy denotes the percentage of documents that are correctly classified based on the keywords. From the Fig. 1 it is evident that proposed classifier ARC is giving the good performance with all the chosen keywords extraction methods in ACM dataset [16], where the dataset holds 3506 documents. It is to be noted that classification algorithms combined with CSI [20] are giving a very low performance in terms of accuracy towards classifying the documents. The result shows that ARC (proposed) gives the top level accuracy in all keyword extraction methods EB [21], MF [22], TF-ISF [23], and specifically with TR [24]. This is due to the segregating the documents in an indiscriminate manner and performing the classification, where RF and Bagging RF makes classification in a sequential manner. Also, RF [18] and Bagging RF [19] classifier was proposed to support the documents which are fully text-based, where ARC is proposed to classify the documents even though multimedia (i.e., image) contents are present.

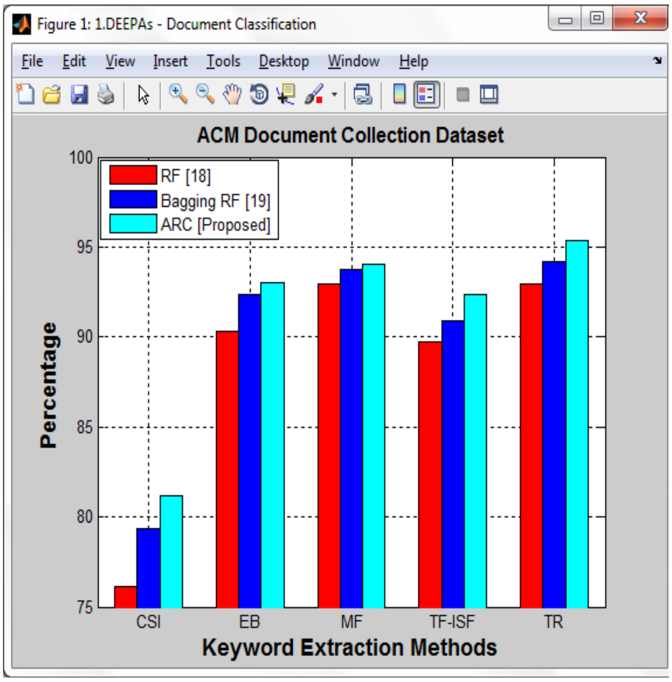


Figure. 1 Classification Accuracy vs ACM Document Collection Dataset

Fig. 2 shows evaluation result of classification algorithms with chosen keyword extraction methods on Reuters-21578 [17] dataset, where the dataset holds 21578 documents. It is clear that ARC classifier is giving better accuracy when combined with keywords extraction methods, and it to be noted that all classification algorithms combined with CSI [20] is giving low accuracy when compared with other methods namely EB [21], MF [22], TF-ISF [23], and TR [24]. It is evident that ARC is giving better classification result when combined with TR [24], this is due to setting the threshold value for processing the documents. ARC does not take all the documents at once and process for classification, instead it sets the threshold value and takes the documents for processing in a batch. Due to this, the classification accuracy is increased. The results shows that existing classification algorithms RF [18] and Bagging RF [19] are not fit for huge dataset like reuters-21578 [17], the time taken for classifying gets delayed too much due to following the sequential manner classification.

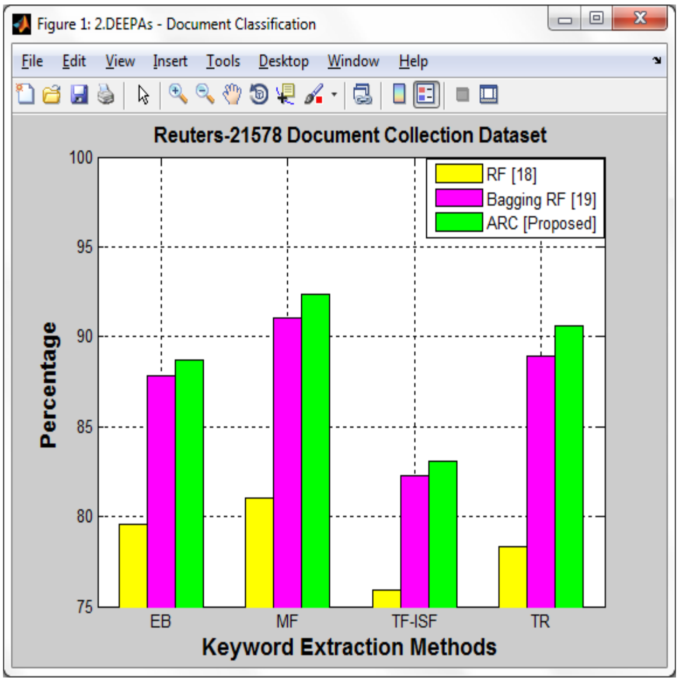


Figure. 2 Classification Accuracy vs Reuters-21578 Document Collection Dataset

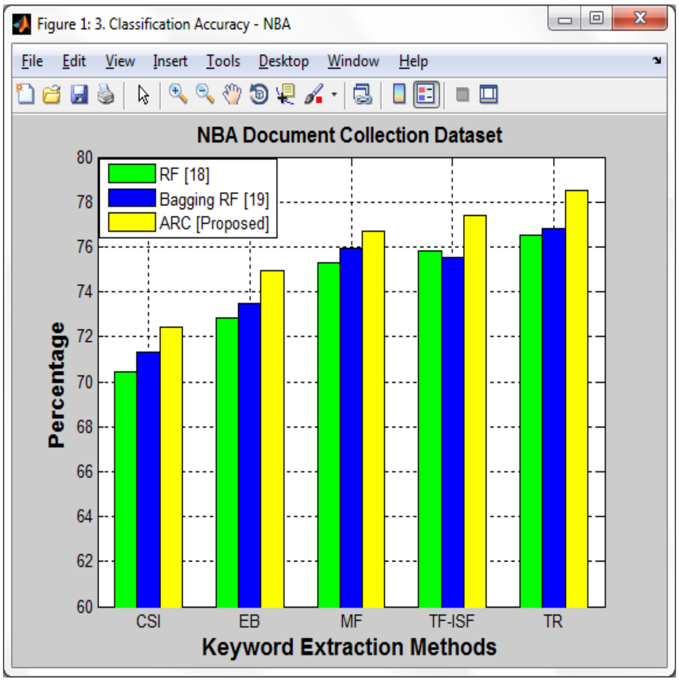


Figure. 3 Classification Accuracy vs NBA Document Collection Dataset

Fig. 3 shows evaluation result of classification algorithms with chosen keyword extraction methods on NBA Input Document Collection Dataset, where the dataset holds 1256 documents. It is clear to see that that the proposed classifier is giving the better accuracy when it is combined with keywords extraction methods. Also RF-CFI [18] gives the poor accuracy due to making the classification in one to one manner.

It is evident that ARC is giving better classification result when combined with TR [24], this is due to utilizing the threshold value concept for processing the documents. Because of using the threshold value, the documents are clustered in a random manner with appropriate groups. Hence the classification accuracy gets increased due to this. The results shows that existing classification algorithms RF [18] and Bagging RF [19] are not fit for NBA Input Document Collection Dataset, the time taken for classifying gets delayed too much due to following the sequential manner classification.

6.2 F-Measure Analysis

F-Measure denotes the percentage of harmonic mean of recall and precision. From Fig. 4 it is noticeable that proposed classifier ARC haven given remarkable performance with ACM dataset [16], where the dataset holds 3506 documents. The classification algorithms combined with CSI [20] have very low F-Measure. The result shows that the proposed classifier ARC is able to give best performance with all the keyword extraction methods (EB [21], MF [22], TF-ISF [23], TR [24]), where RF [18] and Bagging RF [19] classifiers have given the low performance when comparing with the proposed classifier. The proposed classifier ARC is having the best performance in precision and recall, where both are used in the calculation of F-Measure. F-Measure of the proposed classier indicates that ARC is better result in terms of precision and recall.

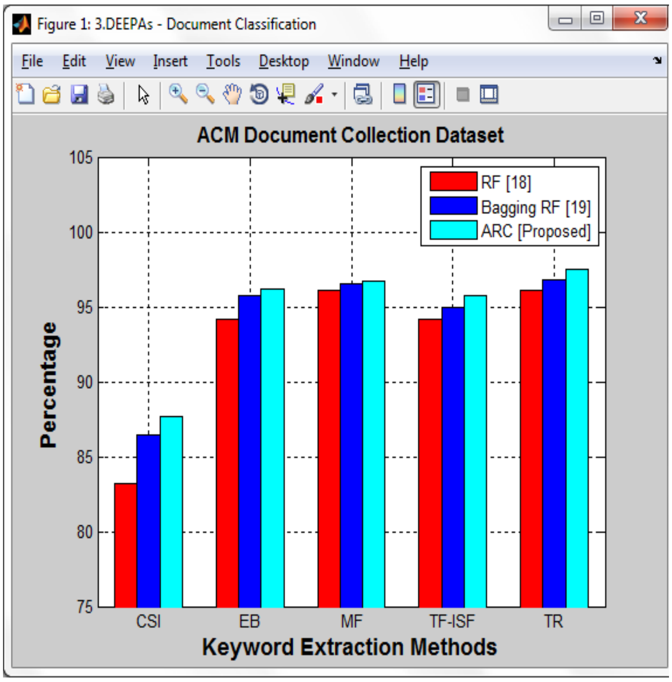


Figure. 4 F-Measure vs ACM Document Collection Dataset

Fig. 5 highlights the F-Measure results of classification algorithms RF [18] and Bagging RF [19] and the proposed classifier ARC. Fig. 5 clearly demonstrate that the proposed classifier ARC have given the better F-Measure result with all keyword extraction methods (CSI [20], EB [21], MF [22], TF-ISF [23], and specifically with TR [24]), where it has given the low F-Measure with CSI [20]. When analyzing the results, it is found that the keywords extracted by CSI [20] is not sufficient to classify the documents, but ARC is able to give highest F-Measure when comparing with RF [18] and Bagging RF [19]. The reason for the best outcome of F-Measure by ARC is it does not take all the documents at once processing the classification, instead it sets the threshold value and takes the documents for processing in a batch. The results shows that ARC is best suitable for huge dataset like Reuters-21578 [17], when it is combined with TR [24].

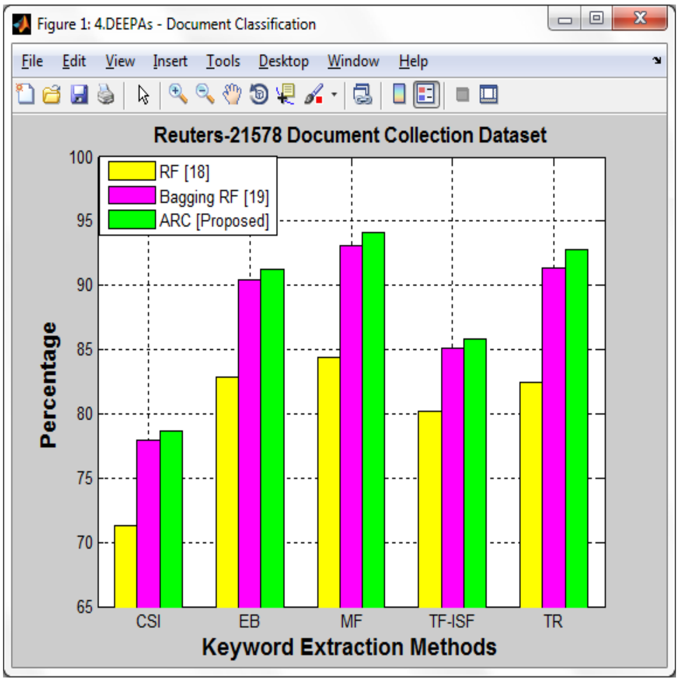


Figure. 5 F-Measure vs Reuters-21578 Document Collection Dataset

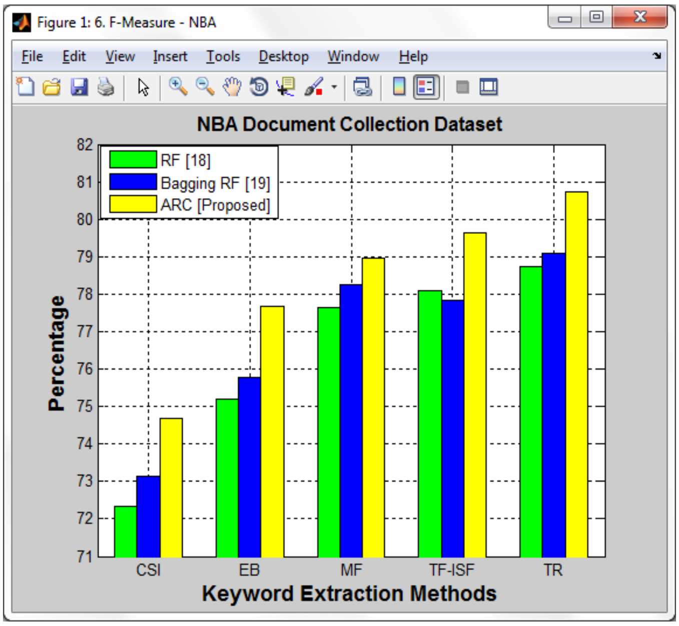


Figure. 6 F-Measure vs NBA Document Collection Dataset

Fig. 6 highlights the F-Measure results of classification algorithms RF [18] and Bagging RF [19] and the proposed classifier ARC. Fig. 6 clearly illustrate that the proposed

Classifier ARC having the better F-Measure result with all keyword extraction methods (CSI [20], EB [21], MF [22], TF-ISF [23], and specifically with TR [24]), where it has given the low F-Measure with CSI [20]. When analyzing the results, it is found that the keywords extracted by CSI [20] is not sufficient to classify the documents, but ARC is able to give highest F-Measure when comparing with RF [18] and Bagging RF [19]. The reason for the best outcome of F-Measure by ARC is, it does not take all the documents at once processing the classification, instead it sets the threshold value and takes the documents for processing in a batch. The results shows that ARC is best suitable NBA Input Document Collection Dataset, when it is combined with TR [24].

VIII. CONCLUSION

This paper has proposed an adaptive robust classifier to classify the documents with more accuracy based on the keywords extracted by different methods. Most available classifiers are suitable for more small or specific dataset. Those classifiers won't have better performance with huge dataset. The proposed classifier is designed to adopt dataset of any size and perform classification by segregating the dataset into multiple parts and perform classification in a random manner which results in an improved classification accuracy and f-measure. For evaluating the performance of the proposed classifier ACM document collection dataset, Reuters-21578 document collection dataset, and NBA Input Document Collection Dataset are used.

REFERENCES

1. K. N. Junejo, A. Karim, M. T. Hassan, M. Jeon, "Terms-based discriminative information space for robust text classification", *Information Sciences*, Volume 372, 2016, Pages 518-538.
2. W. Liang, H. Xie, Y. Rao, R. Y. K. Lau, F. L. Wang, "Universal affective model for Readers' emotion classification over short texts", *Expert Systems with Applications*, Volume 114, 2018, Pages 322-333.
3. F. D. Berdun, M. G. Armentano, L. Berdun, M. Mineo, "Classification of collaborative behavior from free text interactions", *Computers & Electrical Engineering*, Volume 65, 2018, Pages 428-437.
4. J. Hartmann, J. Huppertz, C. Schamp, M. Heitmann, "Comparing automated text classification methods", *International Journal of Research in Marketing*, 2018.
5. T. V. Phan, M. Nakagawa, "Combination of global and local contexts for text/non-text classification in heterogeneous online handwritten documents", *Pattern Recognition*, Volume 51, 2016, Pages 112-124.
6. F. Viegas, L. Rocha, E. Resende, T. Salles, W. Martins, M. Ferreira, E. Freitas, M. André Gonçalves, "Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification" , *Neurocomputing*, Volume 307, 2018, Pages 153-171.
7. X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi, "Text/non-text image classification in the wild with convolutional neural networks", *Pattern Recognition*, Volume 66, 2017, Pages 437-446.
8. C. Tsai, C. Chang, "SVOIS Support Vector Oriented Instance Selection for text classification", Information Systems, Volume 38, Issue 8, 2013, Pages 1070-1083.
9. W. Zhang, F. Gao, "An Improvement to Naive Bayes for Text Classification", *Procedia Engineering*, Volume 15, 2011, Pages 2160-2164.
10. A. Rehman, K. Javed, H. A. Babri, "Feature selection based on a normalized difference measure for text classification", *Information Processing & Management*, Volume 53, Issue 2, 2017, Pages 473-489.
11. F. Ren, M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification", *Information Sciences*, Volume 236, 2013, Pages 109-125.
12. D. Agnihotri, K. Verma, P. Tripathi, "Variable Global Feature Selection Scheme for automatic classification of text documents", *Expert Systems with Applications*, Volume 81, 2017, Pages 268-281.
13. J. Chen, H. Huang, S. Tian, Y. Qu, "Feature selection for text classification with Naïve Bayes", *Expert Systems with Applications*, Volume 36, Issue 3, Part 1, 2009, Pages 5432-5435.
14. A. Rehman, K. Javed, H. A. Babri, M. N. Asim, "Selection of the most relevant terms based on a max-min ratio metric for text classification", *Expert Systems with Applications*, Volume 114, 2018, Pages 78-96.
15. J. L. O. Hui, G. K. Hoon, W. M. N. W. Zainon, "Effects of Word Class and Text Position in Sentiment-based News Classification", *Procedia Computer Science*, Volume 124, 2017, Pages 77-85.
16. R. G. Rossi, R. M. Marcacini, S. O. Rezende, Analysis of Domain Independent Statistical Keyword Extraction Methods for Incremental Clustering, *Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence*, Vol. 12, Iss. 1, 2014, Pages 17-37.
17. A. K. Uysal, "An Improved Global Feature Selection Scheme for Text Classification", *Expert Systems with Applications*, Vol. 43, 2016, Pages 82-92.
18. L. Breiman, Random Forests, *Machine Learning*, Vol. 45, Issue. 1, 2001, Pages 5-32
19. A. Onan, S. Korukoglu, H. Bulut, "Ensemble of Keyword Extraction Methods and Classifiers in Text Classification", *Expert Systems with Applications*, Vol. 57, 2016, Pages 232-247.
20. Y. Matsuo, M. Ishizuka, "Keyword Extraction from a Single Document using Word Co-Occurrence Statistical Information", *International Journal on Artificial Intelligence Tools*, Vol. 13, Issue. 1, pages. 157–169, 2004.
21. G. K. Palshikar, "Keyword Extraction from a Single Document Using Centrality Measures", In: Proc. *Second International Conference on Pattern Recognition and Machine Intelligence*, India. *Lecture Notes in Computer Science*, Vol 4815, pages 503-510, 2007.
22. R. G. Rossi, R. M. Maracini, S. O. Rezende, "Analysis of Domain Independent Statistical Keyword Extraction Methods for Incremental Clustering", *Learning and Nonlinear Models*, Vol. 12, Issue. 1, pages 17–37, 2014.
23. L. J. Neto, A . D. Santos, C. A .Kaestner, A. A. Freitas, "Document Clustering and Text Summarization", In: Proc. *4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, United Kingdom, pp. 41–55, 2000.
24. R. Mihalcea, P. Tarau, "TextRank: Bringing Order into Text", In: Proc. *2004 Conference on Empirical Methods in Natural Language Processing*, Spain, pp. 404–411, 2004.