iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6116

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Analysis of Feature Selection Methods for Text Classification using Multiple Datasets

Archit Aggarwal[1], Bhavya Gola[2], Tushar Sankla[3]
[1, 2, 3]*Department of Information Technology, Bharati Vidyapeeth's College of Engineering, New Delhi*

Abstract: ***Text Classification is the process of accommodating different categories of text on the basis of the content. It is a fundamental task of Natural Language Processing (NLP) having varied applications like sentiment analysis, spam detection, topic labelling and intent labelling. The first step of the classifiers is extraction i.e. to convert words and phrases into vectors which refers to the frequency of a word in a predefined dictionary of words. There are various machine learning algorithms that can be used for classification. In this paper, we will implement best first, information gain and gain ratio feature selection on certain classifiers such as Naive Bayes, Bagging, Random Forest and Naive Bayes Multinomial. We will find and compare the Accuracy, Training Time, Testing Time, Mean Absolute Error and Recall for the feature selections for each classifier. It will help to find which classifier and feature selection method is best suited for performing text classification.***

Index Words: Naive Bayes(NB), Naive Bayes Multinomial(MN), Information Gain(IG), Gain Ratio(GR),Gini Index(GI), Odds Ratio(OR),Chi-Square(CHI),Term Frequency(TF), Document Frequency(DF) Distinguishing feature selector (DFS), Area Under Curve (AUC), Mean absolute error (MAE), Natural Lan- guage Processing(NLP), Machine Learning(ML),Bag of Words (BOW), Customer Relationship Management(CRM)

I. INTRODUCTION

Unstructured data is present in various forms such as chats, emails, web pages, social media, survey responses, support tickets etc. This data can be very useful in providing insights and better decision making. But it can be very hard and time consuming to get insights from unstructured data. So, to make sense of the humongous data, businesses are using text classifiers to organise data, automate the process and to make better decisions. Example, organizing articles by topics, conversations by language, support tickets by urgency, brand mentions by sentiments, chat etc. [6] The economic value of the digital world has increased tremendously because new electronic documents, finding information on the web and guiding the user through hypertext have been possible because of the text classification techniques.

[19]There are two ways for the classification of text: man- ual and automatic classification. The former is time-consuming but the result quality is good as it is done by a human. The latter refers to ML, NLP and other techniques that are faster and cost effective than the former. Text classification using machine learning is done on the basis of the past observations. [16]A ML algorithm uses pre-labeled examples as training data. It learns the association between the pieces of text and that a certain output is required of certain input.

II. RELATED WORK

In this section we review papers and other works related to our research. There are very few studies on assessing feature selection methods for text classification, we are examining performance metrics for text classification from other fields. The method of looking for the minimum size of the correct text attribute is known as feature selection in text classification. There are different methods of classifying text, such as filters, global methods, and local methods. Filters is one of the three preferred texts for categorization. A single score is assigned to a feature in global method regardless of the number of classes. Several scores are given when using the local process, since each element has a score in each class. Algorithms like DF and GI are extensively used for global feature selection. While for local feature selection, algorithms like CHI, OR and the selector DFS can be used.In order to evaluate the feature selection methods, we choose a variety of descriptive classification performance measures of the three groups. In [5] the linear relationship between the different classes of the evaluation measures is limited, but greater in same class. The use of one measure alone does not precisely indicate the effectiveness of a feature selection method. Classification performance is the most common metric for determining the efficiency of selecting a function. Experimental analysis is a good way to evaluate the classification algorithms. Metrics such as accuracy, F1 score, consistency, recall, specificity, AUC, MAE, and mean square error (MSE) can be measured to conduct a quality evaluation.

[17] Existing models have used an enhanced NB Model exploring various alternatives to improve accuracy and recall of the above mentioned classifier for sentiment analysis.It was achieved by the experiment of complications of regression training and time checking. The paper argues that to improve speed and accuracy, the suggested method may be generalized to various text sorting problems. [15] The difference between the multinomial model and the multivariate Bernoulli model has been clarified in another research which helps to explain the paradigm which vocabulary sizes fits best. The paper con- cluded that with a larger vocabulary size, the multinomial model performs well, and the multivariate Bernoulli model performs well in small vocabulary size. The analysis found that there was less error rate on the multinomial test compared to the multivariate Bernoulli model. The assessment criteria used here were accuracy criterion, the time available for preparation and the capital base. [9] It focuses on sentiment analysis and they did a differential analysis based on different parameters to analyse and evaluate the differences between classifiers. [1]Focuses on showing the decline in performance of using traditional Random Forests in the classification of short text compared to using them for standard text. A new approach to improving the performance of text categorization was proposed in the research paper, i.e. to combine data enrichment with the introduction of semantics in Random Forest. This resulted in an improved accuracy compared to the traditional method used earlier.

III. RESEARCH METHODOLOGY

We studied the various steps that are involved in text classification. Then, we look at the methods that convert a cleaned sequence of words to numerical feature vectors.TF- IDF evaluates how relevant a word is to a document in a collection of documents. We have also used feature selectors in our experiment. The ML algorithm has two major steps:

1) *Feature Extraction*: It refers to the transformation of text to vector which represents the frequency of words in the preset dictionary of words. Training data consists of feature sets (vectors) and tags (sports, GK). These training sets are fed in the ML algorithm so as to produce a classification model. [12]
2) *Prediction*: The model is trained with enough training samples so that it can begin making accurate predictions.

A. *Text Preprocessing*

Text preprocessing is necessary for converting text from human language to case-sensitive machine understandable format for further processing. It is a significant step to clean up the data before we commence the classification. The process of text preprocessing consists of : Normalization, Tokenization and Lemmatization.

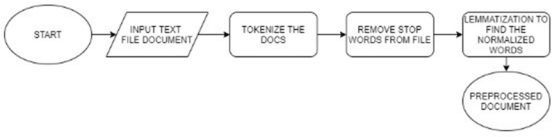


Fig 1: Flowchart of document preprocessing

1) *Normalization:* Normalization involves a series of steps. First we convert all letters to lowercase ,then we convert numbers into words or remove them altogether from the document. Then we remove the punctuations, white spaces expand all the abbreviations. Then we proceed to remove sparse terms and stop words, and perform canonicalization of text.
2) *Tokenization:* The process of tokenization involves breaking down a document into fragments like words, punctuation marks, numeric digits, etc. We can consider words, numbers, punctuation marks etc. as tokens.
3) *Lemmatization:* In the process of lemmatization we make groups of the various inflected forms of any word for it to be considered as one entity. It brings a background or a setting to the lexicons. It connects all words having indistinguishable meaning to the said word. Lemmatization does morphological research of the words.
4) *Text Representation:* One of the main focus of Information Retrieval (IR) and text mining is 'Text Representation'. It's objective is to present the unstructured text documents numerically so it is mathematically computable.
5) *Bag of Words:* method to extract features from text documents.NLP can't process words directly so we create BOW. It keeps count of all the most frequently occurring words in the text.
6) *Word2Vec:* creates words embedding and used for making linguistic sense of words. It creates a large vector space with numerous dimensions with each word being assigned a vector and similar words are grouped closer.

B. TF-IDF

[21]A statistics standard we use for term weighing.vChecks for relevance of a word in a document from a set of documents. [13]

1) *TF:* word frequency in a document
2) *IDF:* tells us whether the word is frequent or not. If the word is closer to 0 then it is considered to be frequent and if closer to 1 then rare.

To find TF-IDF score:

Where:

a: word in a document b: document

S: set of documents

TF-IDF(a,b,S) = TF(a,b).IDF(a,S)

where:

TF(a,b) = log(1+freq(a,b))

IDF(a,S) = log(Total no. of docs. / No. of docs. containing term a in it).

The TF-IDF score can be used as input for various algorithms like Naive Bayes and Naive Bayes Multinomial.

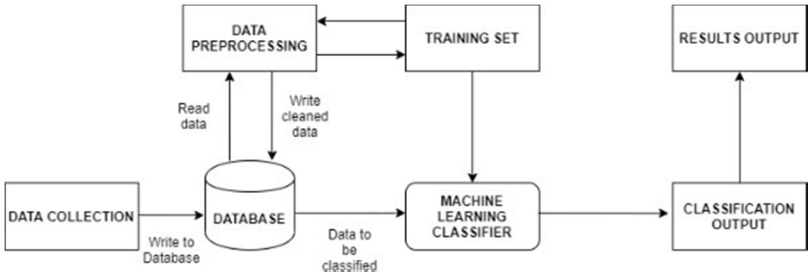


Fig 3. Analysis of the input data

C. Classifiers

For our research, we used Naive Bayes, Bag- ging, Random Forest and Naive Bayes Multinomial classifiers for text classification on various datasets.

1) *Naive Bayes*: It is a very simple probabilistic model that tends to work well on text classification. This model works on simplifying conditional independence assump- tion i.e. the words are conditionally independent of each other irrespective of the given class being positive or negative. [18] [8]
2) *Bagging*: The bagging classifiers use bootstrap sampling of the training data to build n classification trees.The predictions from them are used to make a final meta- prediction. Bagging basically improves the estimate of one by combining the estimates of many. [7]
3) *Random Forest*: It is a supervised learning method for regression and classification techniques. It is very easy and flexible to use. Random decision forests create trees on some arbitrary data samples, then through voting, it selects most suitable solutions from a range of predictions of decision trees. [22]
4) *Naive Bayes Multinomial*: The multinomial classifier is a specialized version of Naive Bayes classifier specifically used for text documents. It explicitly focuses on word count information in documents. This algorithm uses a training classifier with help of available documents and uses probabilistic labels for the documents that are unlabeled. It also trains a classifier that is new using the labels for all the documents and it iterates to convergence. [15]

D. Feature Selection

[3]This process selects only a few attributes from all the instances present in the train dataset based on ranking of their contribution to a class and uses only these attributes as features in text classification while removing rest of the features. The main importance of feature selection is that it can be used to remove either repeating or unnecessary information from a model without compromising any credibility and also increases the precision of classification. It narrows the size of effective attributes and reduces training time.

1) *Information Gain* : It measures the decrease in the entropy by dividing a dataset compatible with said value of stochastic variable. Greater value IG implies a lower entropy sample group and then lesser nonplus. Events with lower probability have more information as the uncertainty is more whereas in events with a greater probability have significantly less to tell i.e less in- formation. Entropy measures the information proportion present in the said stochastic variable, or its distribution of probability. Information Gain= IG(T, a) = H(T) – H(T— a)

2) *Best First* : Best-First selects the n best features for modeling a given dataset, using a greedy algorithm. It starts by creating N models in which each of them uses only one of the N features of our dataset as input. The feature that produces the model with the best performance is selected for further iterations. In the next iteration, it creates another set of N-1 models with two input features: one is selected from the previous iteration and another from the N-1 remaining features. It stops when it gets desired number of features.
3) *Gain Ratio* : Gain ratio is a variation of IG technique that helps in decreasing the bias towards multi valued attributes. It takes into account sizes and number of branches while selecting an attribute. Used for correcting the IG by utilizing the intrinsic information . [11] G(A) = G.R(A)/Intrinsic Info(A) where; G: Gain, G.R: Gain Ratio, A: Attribute

E. Classification Parameters

These parameters are used for ranking on the basis of performance. In this paper we have analysed the model by comparing Mean Absolute Error, Recall, Accuracy, Build Time and Testing Time. The present studies mostly focus on AUC(area under the curve) and accu- racy. Cortes [10] showed that "algorithms designed to minimize the error rate may not lead to the best possible AUC values." Davis and Goadrich [4] found that methods perform well in the receiver operating characteristic (ROC) space only if they perform well in the precision–recall space.

1) *Accuracy*: Accuracy is a commonly used parameter for assessing classifier results. It is the part of recovered documents that are germane to the query.
2) *Recall*: It is also called sensitivity, is the part of the pertinent documents that are successfully recovered. If recall is high it means lesser false negatives, while having low value of recall implies greater false negatives.
3) *Mean Absolute Error* : MAE is a parameter for taking the average of all absolute errors.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$ [20]

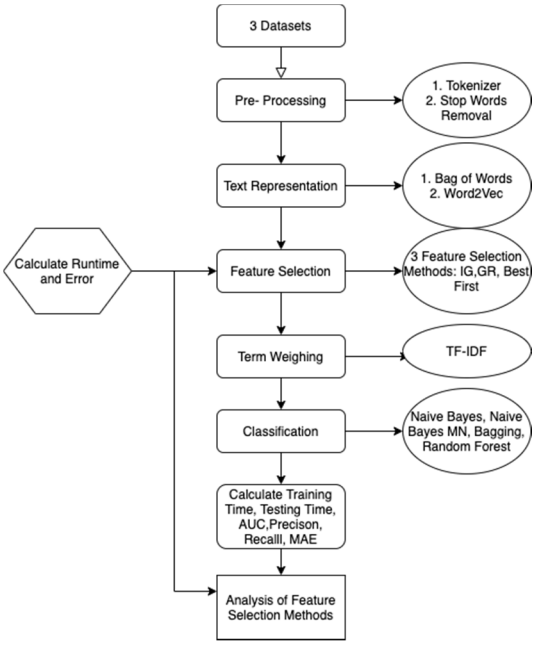
4) *Time Taken(Build)* : It depicts the time consumed by the classifier to build the model for a given dataset. Every algorithm wants to achieve time complexity as low as possible. So it is of utmost value as it makes the comparison more constructive and reliable.
5) *Time Taken(Test)* : It depicts the time taken for testing the results by the algorithm for a given dataset.

F. Datasets Used

1) *20Newgroups* :The dataset consists two sub-datasets: one to train the model and the other to test the perfor- mance .The division between the training set and testing set is established on conversations reported prior to and following a particular time period having 20 topics with several thousand posts [14]
2) *Polarity Dataset* : Polarity dataset contains two columns namely pos and neg resembling positive and negative reviews. The positivity negativity were deter- mined by checking if the reviews were above 5 for a 10 point system.
3) *Reuters21578* : Similar to the newsgroup data set it has news articles from reuters news wire. It contains 10,788 documents which has 2 subsets: a training set with 7769 documents and a test set with 3019 documents. [2]

IV. METHODOLOGY

In this paper our main objective is to calculate the run time and accuracy. We input data from the three datasets that we have mentioned in section 3.6 i.e. 20 Newsgroup, Polarity dataset and Reuters 21578 dataset. Firstly we pre-process them as mentioned in section 3.1. We perform normalization, tokenization and lemmatization. The main objective of text preprocessing is to convert text into machine understandable format. The next step focuses on the representation of the text. The two processes for it are BOW and Word2Vec. BoW trains the machine learning algorithms and Word2Vec converts word to vectors i.e represents word features as numbers (0 or 1). It does so without human interference. After text preprocessing and text representation we perform the various feature selection methods as mentioned in section3.4. In this paper we have implemented Information Gain, Best First and Gain Ratio. We perform different feature selections on a classifier to find out which selection gives us better accuracy and less error rate. The next step is term weighting which scores words in ML algorithms for NLP. This we have done by TF-IDF, (section 3.2). Here, we have performed feature selection on four classifiers: Naive Bayes, Random Forest, Bagging and Naive Bayes Multinomial Classifier (Section 3.3) They help us to compare the various evalua- tion metrics such as Mean Absolute Error, Accuracy, Recall, Build Time and Testing Time as mentioned in section 3.5.



V. RESULT

We have compiled various tables for all the datasets comparing the values of the aforementioned classification parameters with and without feature selection for each of the 4 classifiers individually which are as follows : Naive Bayes, Naive Bayes Multinomial, Random Forest and Bagging.

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.0156	0.0152	0.0141	0.0143
Accuracy	0.987	0.987	0.987	0.987
Recall	0.987	0.987	0.987	0.987
Time taken(build)	24.4 s	0.39 s	24.47 s	23.07 s
Time taken(test)	0.04 s	0.02 s	0.04 s	0.04 s

TABLE I
BAGGING CLASSIFIER ON REUTERS DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.0704	0.0154	0.0704	0.0704
Accuracy	0.978	0.991	0.978	0.978
Recall	0.930	0.989	0.930	0.930
Time taken(build)	0.95 s	0.04 s	1.14 s	0.85 s
Time taken(test)	1 s	0.02 s	0.92 s	0.88 s

TABLE II
NAIVE BAYES CLASSIFIER ON REUTERS DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.0128	0.0202	0.0128	0.0128
Accuracy	0.992	0.996	0.992	0.992
Recall	0.989	0.996	0.989	0.989
Time taken(build)	0.06 s	0 s	0.02 s	0.01 s
Time taken(test)	0.07 s	0 s	0.04 s	0.03 s

TABLE III
NAIVE BAYES MULTINOMIAL CLASSIFIER ON REUTERS DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.0355	0.0111	0.0334	0.0337
Accuracy	0.977	0.994	0.977	0.977
Recall	0.977	0.994	0.977	0.977
Time taken(build)	6.06 s	0.69 s	5.57 s	5.78 s
Time taken(test)	0.31 s	0.02 s	0.12 s	0.14 s

TABLE IV
RANDOM FOREST CLASSIFIER ON REUTERS DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.4473	0.3387	0.4412	0.4435
Accuracy	0.742	0.773	0.756	0.761
Recall	0.740	0.771	0.754	0.760
Time taken(build)	8.57 s	3.61 s	8.97 s	9.12 s
Time taken(test)	0.21 s	0.14 s	0.2 s	0.23 s

TABLE V
RANDOM FOREST CLASSIFIER ON POLARITY DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.2101	0.2203	0.2101	0.2101
Accuracy	0.791	0.800	0.791	0.791
Recall	0.790	0.800	0.790	0.790
Time taken(build)	1.05 s	0.05 s	0.99 s	1.1 s
Time taken(test)	1.32 s	0.04 s	0.58 s	0.85 s

TABLE VI
NAIVE BAYES CLASSIFIER ON POLARITY DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.2052	0.2642	0.2052	0.2052
Accuracy	0.807	0.799	0.807	0.807
Recall	0.807	0.799	0.807	0.807
Time taken(build)	0.03 s	0	0.03 s	0.03 s
Time taken(test)	0.04 s	0	0.05 s	0.04 s

TABLE VII
NAIVE BAYES MULTINOMIAL CLASSIFIER ON POLARITY DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.4137	0.3623	0.4118	0.4119
Accuracy	0.666	0.736	0.666	0.666
Recall	0.666	0.735	0.666	0.666
Time taken(build)	56.8 s	1.89 s	64.31 s	58.08 s
Time taken(test)	0.05 s	0.01 second	0.05 second	0.05 s

TABLE VIII
BAGGING CLASSIFIER ON POLARITY DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.0315	0.0527	0.0527	0.0315
Accuracy	0.732	0.569	0.569	0.732
Recall	0.686	0.514	0.514	0.686
Time taken(build)	57.83 s	0.18 s	0.25 s	41.25 s
Time taken(test)	172.26 s	1.59 s	3.29 s	154.36 s

TABLE IX
NAIVE BAYES CLASSIFIER ON NEWSGROUP DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.0805	0.0673	0.0672	0.0804
Accuracy	0.718	0.493	0.495	0.730
Recall	0.708	0.489	0.490	0.714
Time taken(build)	324.7 sec	27.27 s	42.12 s	432.94 s
Time taken(test)	7.23 s	0.96 s	1.54 s	2.95 s

TABLE X
RANDOM FOREST CLASSIFIER ON NEWSGROUP DATASET

	No feature Selection	Best First	Information Gain	Gain Ratio
Mean Absolute Error	0.016	0.0581	0.0581	0.016
Accuracy	0.853	0.547	0.547	0.853
Recall	0.848	0.532	0.532	0.848
Time taken(build)	0.08 second	0 s	0.1 s	0.12 s
Time taken(test)	0.66 s	0.03 s	0.15 s	0.99 s

TABLE XI
NAIVE BAYES MULTINOMIAL CLASSIFIER ON NEWSGROUP DATASET

VI. CONCLUSION

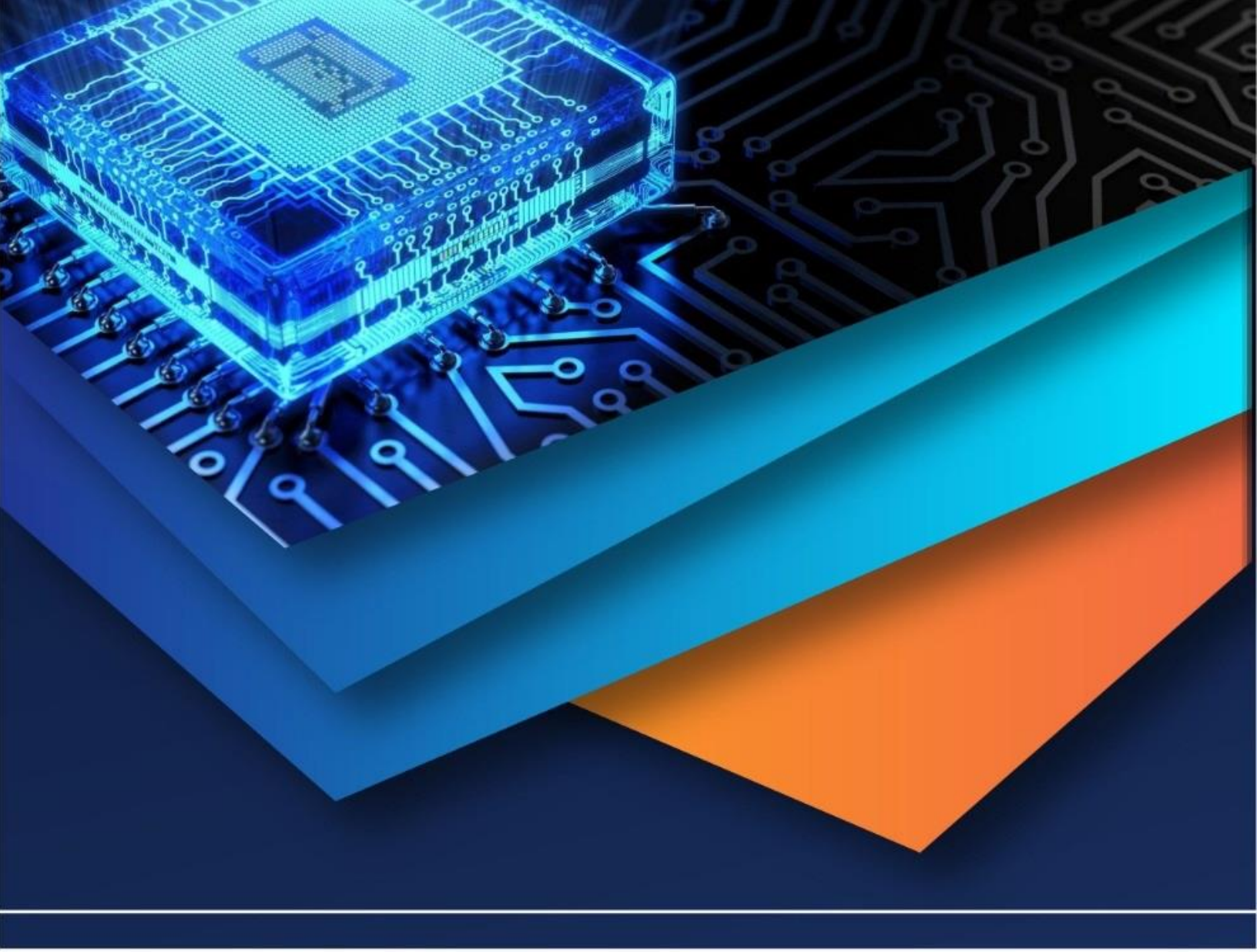
In the study conducted we have performed text classification on 3 different data-sets using three feature selection methods, using 4 text classifiers to compare all the feature selection methods with results from same classifiers without doing feature selection. The main problem that we encounter is to find which feature selection is the best since we have multiple criteria for comparison and does it improve the results as com- pared to not using any feature selection. We can compare the feature selection on the basis of any one particular evaluation metric and then compare the performances. For example, we can compare the Accuracy for all the classifiers for any one feature selection. Naive Bayes Multinomial Classifier has the best accuracy in general among all classifiers with or without using any feature selection and when it comes to handling big data-sets like 20newsgroup or Reuters it takes the least time to train and test the data set to give results.It usually gives more often than not, better or equally good results without using any feature selection as compared to using feature selectors taking into account all the evaluation measures.

VII. FUTURE SCOPE

Apart from normal feature selection methods like Best First, Information Gain, Chi Square etc. There are various other optimization techniques like Genetic Algorithms and Swarm Optimization Techniques for e.g - Artificial Bee Colony(ABC), Firefly Algorithm, Ant Colony Optimization(ACO) which can be used to give better optimized results. Various fields such as marketing, governance and product management are already utilising text classification. E-Commerce platforms, news agencies etc. can use text classification to their benefit so as to make the user experience better by improving on content classification and tags. It can also automate Customer Relationship Management(CRM). Text classifiers help with CRM tasks to be directly analysed and assigned based on the relevance and importance. It plays a significant role in Search Engine Optimization(SEO). And with classifiers with less error rate and more accuracy, it helps in research and an- alyzing tags much more efficient. Academia, law researchers, non-profit organizations etc. encounter a lot of unstructured data, but handling data becomes much easier with tags and categorization.

REFERENCES

[1] Bouaziz, A., Dartigues-Pallez, C., da Costa Pereira, C., Precioso, F., Lloret, P.: Short text classification using semantic random forest. In: In- ternational Conference on Data Warehousing and Knowledge Discovery, pp. 288–299. Springer (2014)
[2] Dalmau, M.C., Flórez, O.W.M.: Experimental results of the signal processing approach to distributional clustering of terms on reuters- 21578 collection. In: European Conference on Information Retrieval, pp. 678–681. Springer (2007)
[3] Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., Mahoney, M.W.: Feature selection methods for text classification. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 230–239 (2007)
[4] Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning, pp. 233–240 (2006)
[5] Ferri, C., Hernández-Orallo, J., Modroiu, R.: An experimental compar- ison of performance measures for classification. Pattern Recognition Letters **30**(1), 27–38 (2009)
[6] Fragos, K., Maistros, Y., Skourlas, C.: A weighted maximum entropy language model for text classification. In: NLUCS, pp. 55–67 (2005)
[7] Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regres- sion: a statistical view of boosting (with discussion and a rejoin- der by the authors). Ann. Statist. **28**(2), 337–407 (2000). DOI 10.1214/aos/1016218223. URL https://doi.org/10.1214/aos/1016218223
[8] Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine learning **29**(2-3), 131–163 (1997)
[9] Gupte, A., Joshi, S., Gadgul, P., Kadam, A., Gupte, A.: Comparative study of classification algorithms used in sentiment analysis. Interna- tional Journal of Computer Science and Information Technologies **5**(5), 6261–6264 (2014)
[10] Han, G., Zhao, C.: Auc maximization linear classifier based on active learning and its application. Neurocomputing **73**(7-9), 1272–1280 (2010)
[11] Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques, waltham, ma. Morgan Kaufman Publishers **10**, 978–1 (2012)
[12] Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science &Business Media (2009)
[13] Jabbari, S., Allison, B., Guthrie, D., Guthrie, L.: Towards the orwellian nightmare: separation of business and personal emails. In: Proceedings of the COLING/ACL on Main conference poster sessions, pp. 407–411. Association for Computational Linguistics (2006)
[14] Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., Alsaadi, F.E.: Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Applied Soft Comput- ing **86**, 105,836 (2020)
[15] McCallum, A., Nigam, K., et al.: A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol. 752, pp. 41–48. Citeseer (1998)
[16] Mitchell, T.M., et al.: Machine learning (1997)
[17] Narayanan, V., Arora, I., Bhatia, A.: Fast and accurate sentiment classification using an enhanced naive bayes model. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 194–201. Springer (2013)
[18] Narayanan, V., Arora, I., Bhatia, A.: Fast and accurate sentiment classification using an enhanced naive bayes model. In: H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li, X. Yao (eds.) Intelligent Data Engineering and Automated Learning – IDEAL 2013, pp. 194–201. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
[19] Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. In: Learning for Text Categoriza- tion: Papers from the 1998 workshop, vol. 62, pp. 98–105. Madison, Wisconsin (1998)
[20] Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research **30**(1), 79–82 (2005)
[21] Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) **26**(3), 1–37 (2008)
[22] Xu, B., Guo, X., Ye, Y., Cheng, J.: An improved random forest classifier for text categorization. JCP **7**(12), 2913–2920 (2012)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY