International Journal on Emerging Technologies **10**(1): 148-153(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Comparative Study on Performance of Document Classification Using Supervised Machine Learning Algorithms: KNIME

Syed Muzamil Basha[1], K. Bagyalakshmi[2], C. Ramesh[3], Robbi Rahim[4], R. Manikandan[5] and Ambeshwar Kumar[5]

[1]Assistant Professor, SKCET, Coimbatore, (Tamilnadu), INDIA
[2]Sri Ranganathar Institute of Engineering and Technology, Coimbatore (Tamilnadu), INDIA
[3]Associate Professor, Department of Computer Science & Engineering, Bannari Amman Institute of Technology, Sathyamangalam, (Tamilnadu), India
[4]Department of Management, Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia
[5]School of Computing, SASTRA Deemed University (Tamilnadu), INDIA

(Corresponding author: R. Manikandan)
(Received 02 March 2019, Revised 31 May 2019 Accepted 10 June 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In digitalizing the data, Document classification needs to perform for effective huge data organization that saves a lot of user time and helps in analyzing customer feedback. Machine Learning Algorithm (MLA) can be used in designing such new logic that can classify the document and manage it when anew instance of data arrives, without human intervention. The objective of our research is to discover the best MLA in Document classification using Document Term Vector representation method. In previous work, the authors are used Document similarity index and cosine measure in classifying the documents. Whereas in the present work, Term frequency is used as a feature in Document Classification. A new freely available Analytical platform named KNIME is used for the experiment. Where, two data sets namely Human-Aids and Mouse Cancer with 150 instances each are considered, with evaluation parameters Recall, Precision and F-score. The finding in our research is that Support Vector Machine (SVM) with 98.889% perform better than Decision Tree (DT) with 96.667% followed by K Nearest Neighbor (KNN) with 84.444% in terms of classification accuracy. In future, we would like work on different datasets using the designed workflow for validation.

Keywords: Document classification, Machine Learning Algorithm, KNIME, Support Vector Machine, Decision Tree, K Nearest Neighbor, Recall, Precision and F-score.

I. INTRODUCTION

As Machine Learning [1] focus on updating the program logic, that can help machine in self Learning when it is exposed to anew instance of data. There are two ways by which machine any perform learning, which is supervised and unsupervised learning [2]. In supervised learning, the training data have both input data and the target value. Whereas in, unsupervised learning only input data is provided to the machine. To solve classification problem using MLA like SVM [3], DT [4] and KNN [5], one should use supervised learning as the data have both input values and response value. The objective of our research is to find the best classification algorithm used for document classification that helps in the effective organization of huge data in documents. In our experiments, We consider two sets of documents, one with Human aids and the other with Mouse and cancer data, downloaded from PubMed. In which, documents are concatenated and preprocessed using filtering and stemming. Later, transformed into Binary Document Vector [6]. Finally, measure the classification accuracy using SVM, DT, and KNN. The contribution of this research work is to make the readers understand the workflow in Document classification [7] using KNIME. The author in [12] had achieved a classification accuracy of 83% on Wikipedia Database in Document Classification. Whereas, we achieved 98.8% classification accuracy using Term Frequency as a feature in classifying the documents, from this it is sure that instead of Document similarity index measure, Term Frequency measure give as best accuracy in classifying the documents, inspired by the recent process in the area of Document Classification. The contribution of our research is to make the readers understand the workflow easily in solving the document classification problem. The advantage of our research is the result can be easily replicated using KNIME and further extended in order to validate the research carried out in this paper.

This paper is organized as follows: In Introduction, the background knowledge requires to understand the classification problem in discussed along with advantages and limitation of the present research work. In Literature review, the research work carried out in the past is discussed with the impact of the present research work. In Methodology, the work flow designed using KNIME is explained in detail with all the Nodes used in KNIME. In Result, the performance of the MLA's is compared and made discussion on the finding in our research.

II. LITERATURE REVIEW

In [8] the author addresses the method to tune the parameters, used weighted fuzzy logic in assigning weights input data to extract sentiments from the trained data and achieved good F-score. Whereas in [9] the author made a detailed comparison of predictive models and performed analysis on Time series dataset. In [10] the author performed analysis on PIMA diabetes dataset and predicted the levels of diabetes based on insulin feature. Whereas in [11] the author used gradient ascent algorithm in finding out the exact weights of the terms used in determining the sentiment of the tweet and used Boosting approach to improve the accuracy of the linear classifier. In [15], the author made an attempt to develop an recommender system, helping in searching the item, that might out found by themselves, In which precision and recall measures are used in measuring the performance of proposed model. In [16], the author made an research in solving the problem in Diabetic Retinopathy. In which, the author proposed a Model, which can capable of calculating the weights, that gives severity level of the patient's eye by using weighted Fuzzy C-means algorithm. In [17], the author proposed a build a model for airlines, that can performs sentiment analysis on customer feedback and achieved Vital accuracy. Where as in [18], the author experimented on finding out the impact of feature selection on overall sentiment analysis and stated that Term frequency have greater impact on analyzing sentiments rather than bigram approach. In [19], binary classification is performed on Labeled tweets on Matlab using Conventional neural network.

Table 1.

Author	Approach	Advantages	Disadvantages
Wu *et al.* 2017 [12]	Make use of heuristic selection rules in quickly picking out related concepts form any given document	Computing similarity Index among Document submitted	Matching small number of related Topics
Van *et al.* 2017[13]	Used Dirichlet process mixture models	Computed cosine similarity among Document	The only SVM is used in determining the classification accuracy
Conneau *et al.* 2017 [14]	Used Convolution Neural Network of layers about 29.	Used Gradient Decent approach in training the first two layers	Convolution Networks are mainly applicable for images processing

From the Literature review made, it is evident that the work on Document classification using Term Frequency on the dataset considered in our experiments are novel, designed workflow using KNIME helps the researchers to understand the basis of document classification.

A. Methodology

To develop the design using KNIME, First, we use the Table Reader node for importing the data and later combined into one document using concatenate node as shown in Fig. 1. The output from the concatenate node is preprocessed to obtain the clean data from the analysis as shown in Fig. 2. From the preprocessed data Term Frequency is derived using Term Filtering node and the same is visualized in figure 3 for top three instances.

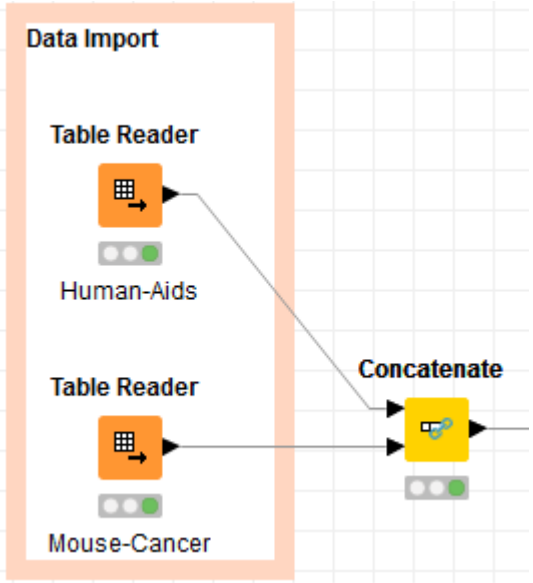


Fig. 1. Data Import Stage.

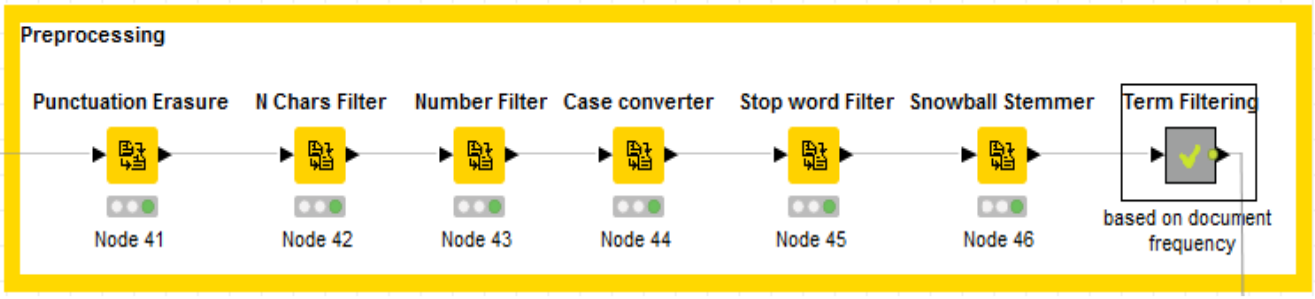


Fig. 2. Preprocessing Stage.

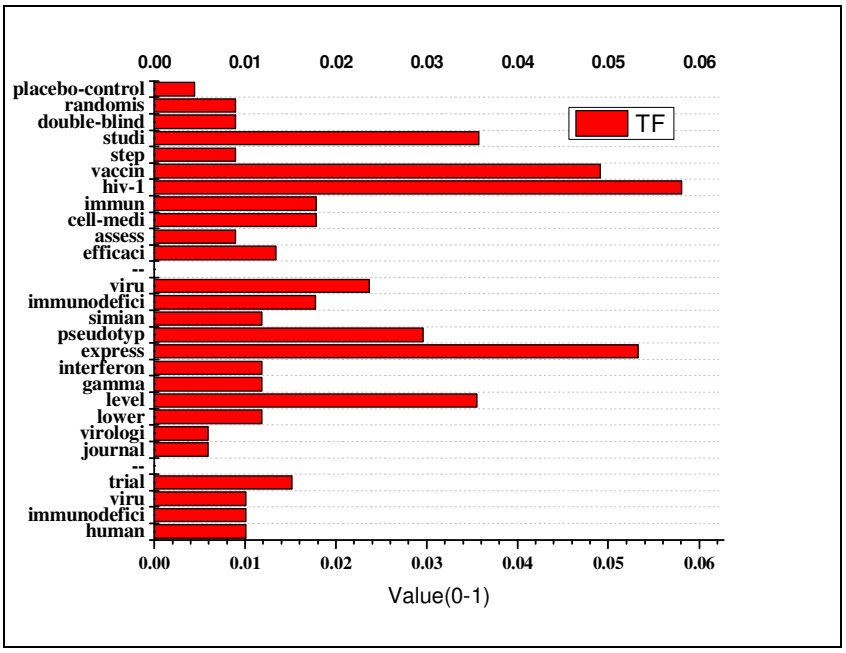


Fig. 3. Term Frequency of Top three instances from both the Dataset.

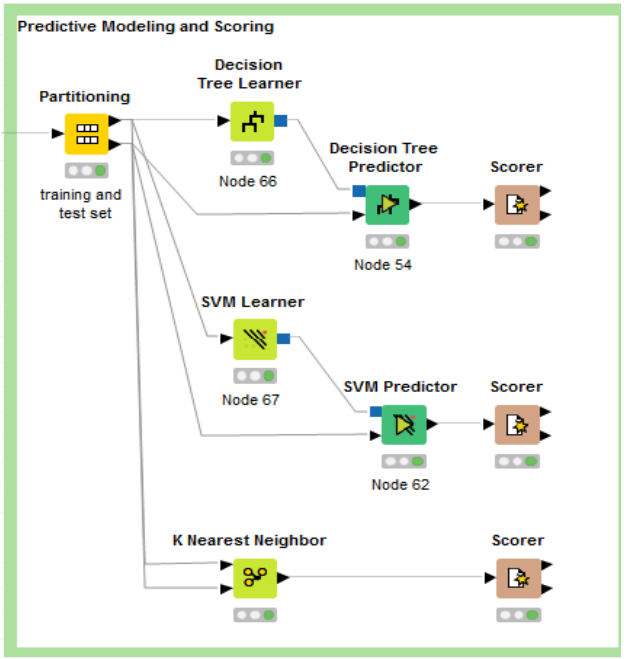


Fig. 4. Predictive Modeling and Scoring.

The result obtained from our experiments are from Scorer node as shown in figure 4. The predictive models used in the experiments are SVM, DT, and KNN.

Support Vector Machine: It works on optimal decision boundary in classifying the object by drawing the orbitary line called Hyper plane based on support vectors by ignoring training samples. It can be used on multidimensionality dataset using polynomial, sigmoid kernels. The best result is obtained by tuning the parameters using K-fold cross-validation. The performance of SVM will be reduced when a number of features aregreater than the number of samples. It can be applied to the area of image interpolation, Healthcare, Financial analysis, pattern recognition.

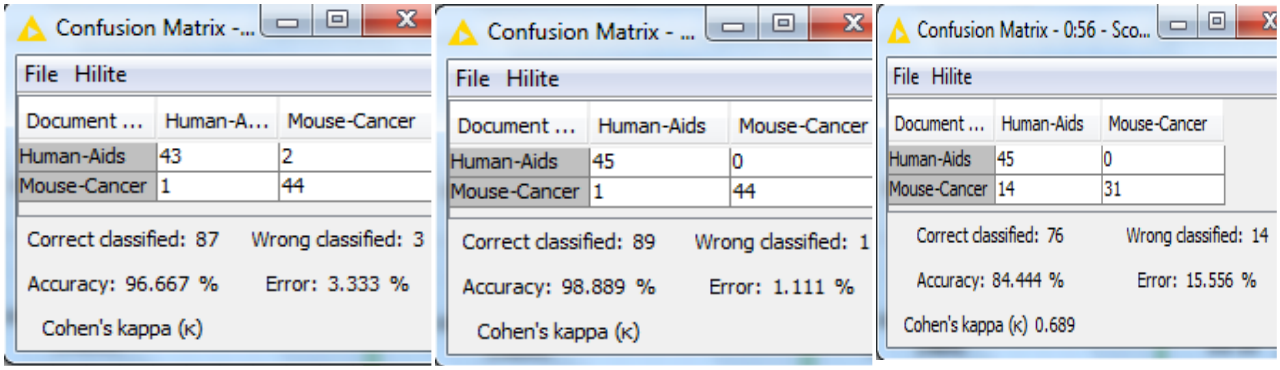
Decision Tree: It is also used for addressing classification problem. In which each node represents an attribute, the Branch represents theoutcome of the set of attribute, and the leaf node represents the class label. It is simple to understand, interpret and visualize, variable screening can be performed both on numerical and categorical data. The drawback of the decision tree is overfitting, a small variation in data resulting in a completely different tree. Bagging and Boosting are used to avoid the variance problem. It is used in the area of Bankruptcy.

K Nearest Neighbor: Mostly Euclidean distance is used to find the near neighbor as in the equation 1. It is used in text categorization, stock market forecasting, pattern recognition, bank customer profiling.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

III. RESULT AND DISCUSSION

In figure 5, the confusion matrix with the accuracy of the MLA used in our experiments is presented. The values of the confusion matrix are used in finding out the evaluation parameters as discussed in Table 2.



Decision Tree SVM KNN

Fig. 5. Confusion Matrix with Accuracy and classification Error.

Table 2: Evaluation Parameters.

Parameters	Formula
Accuracy(A)	$A = \frac{TP + TN}{Total}$
Misclassification Rate(MCR)	$MCR = \frac{FP + FN}{Total}$
Recall(R)	$R = \frac{TP}{Actual\,True}$
Precision(P)	$P = \frac{TP}{\Pr edicted\,True}$
Prevalence(PV)	$PV = \frac{Actual\,True}{Total}$
F Score(FS)	$FS = 2 \times \frac{R \times P}{R + P}$

In figure 6 and 7 the result obtained are plotted for easy comparison of MLA on both the dataset consider in our experiment. Result obtained in our experiments are proved to be the best compared to the work in [12] in terms of classification accuracy, as we consider the Term frequency in classifying the documents. The author in [12] had achieved a classification accuracy of 83% on Wikipedia Database in Document Classification. Whereas, we achieved 98.8% classification accuracy using Term Frequency as a feature in classifying the documents, from this it is sure that instead of Document similarity index measure, Term Frequency measure give as best accuracy in classifying the documents. The contribution of our research is to make the readers understand easily the workflow in solving the document classification problem.

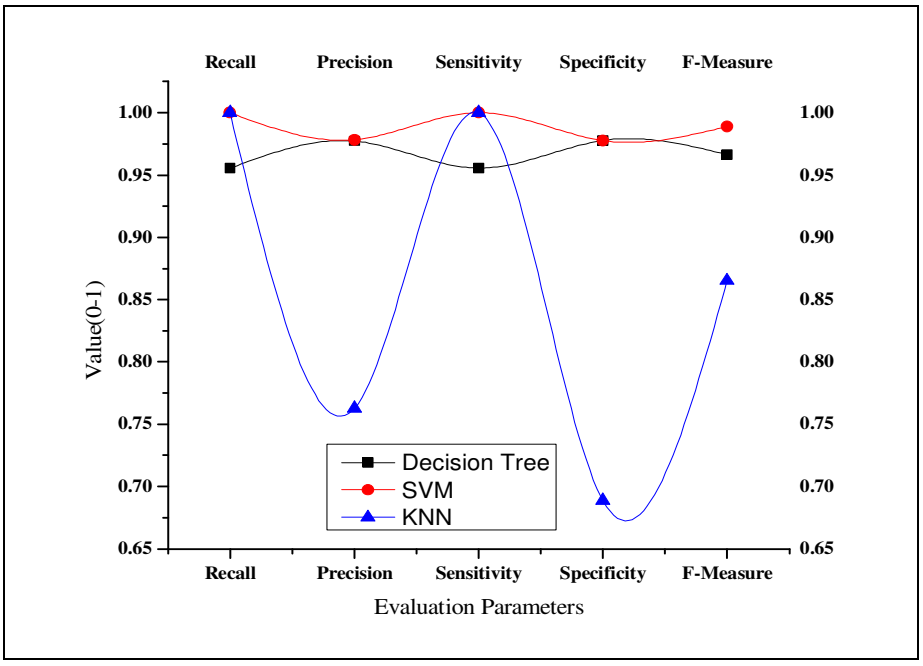


Fig. 6. Score of Machine Learning Algorithm on Human Aids Dataset.

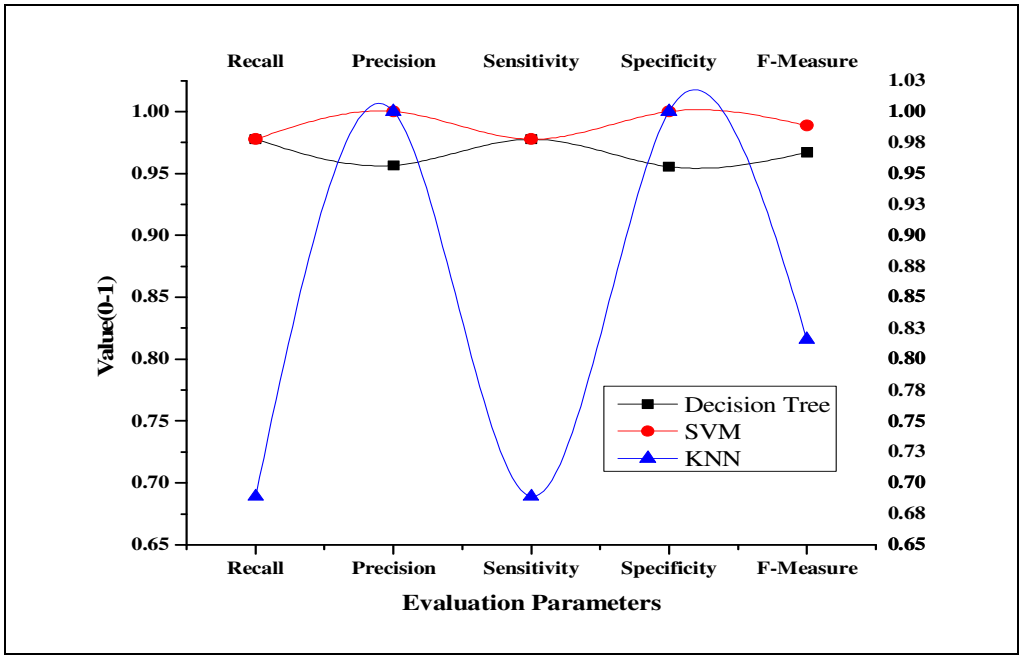


Fig. 7. A score of Machine Learning Algorithm on Mouse Cancer Dataset.

IV. CONCLUSION

Our observation from the present experimental work stats that SVM is the best-supervised machine learning algorithm used in the area of document classification, compared to DT and KNN. The finding in our research is that Support Vector Machine (SVM) with 98.889% perform better than Decision Tree (DT) with 96.667% on both the dataset considered in our experiments followed by K Nearest Neighbor (KNN) with 84.444% in terms of classification accuracy. The challenges in the present research work are to make the data ready for analysis. In future, we would like work on different datasets with more number of instances and apply the machine Learning Algorithms towards better prediction results using the designed workflow for validation on KNIME.

Conflict of Interest: Nil

REFERENCES

[1]. Goldberg, D.E., and Holland, J.H. (1988). Genetic algorithms and machine learning. *Machine learning*, **3**(2): 95-99.

[2]. Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning [book reviews]. *IEEE Transactions on Neural Networks*, **20**(3): 542-542.

[3]. Suykens, J.A., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, **9**(3): 293-300.
[4]. Friedl, M.A., and Brodley, C.E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, **61**(3): 399-409.
[5]. Zhang, H., Berg, A.C., Maire, M., and Malik, J. (2006). SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. **2**, pp. 2126-2136). IEEE.
[6]. Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, **24**(5): 513-523.
[7]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, **11**(1): 10-18.
[8]. Basha, S.M., Zhenning, Y., Rajput, D.S., Iyengar, N., and Caytiles, D.R. (2017). Weighted fuzzy rule based sentiment prediction analysis on tweets. *International Journal of Grid and Distributed Computing*, **10**(6): 41-54.
[9]. Basha, S.M., Zhenning, Y., Rajput, D.S., Caytiles, R. D., and Iyengar, N.C.S. (2017). Comparative study on performance analysis of time series predictive models. *International Journal of Grid and Distributed Computing*, **10**(8): 37-48.
[10]. Basha, S.M., Balaji, H., Iyengar, N.C.S., and Caytiles, R.D. (2017). A Soft Computing Approach to Provide Recommendation on PIMA Diabetes. *International Journal of Advanced Science and Technology*, **106**(1): 19-32.
[11]. Basha, S.M., Rajput, D.S., and Vandhan, V. (2018). Impact of gradient ascent and boosting algorithm in classification. *International Journal of Intelligent Engineering and Systems (IJIES)*, **11**(1): 41-49.
[12]. Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J. and Xu, G. (2017). An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences*, **393**(1): 15-28.
[13]. Van Linh, N., Anh, N.K., Than, K., and Dang, C.N. (2017). An effective and interpretable method for document classification. *Knowledge and Information Systems*, **50**(3): 763-793.
[14]. Conneau, A., Schwenk, H., Barrault, L., and Lecun, Y. (2017). Very deep convolutional networks for text classification. *In Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, **1**(1): 1107-1116.
[15]. Basha, S.M., and Rajput, D.S. (2018). A supervised aspect level sentiment model to predict overall sentiment on tweeter documents. *International Journal of Metadata, Semantics and Ontologies*, **13**(1): 33-41.
[16]. Dutta, S., Manideep, B.C., Basha, S.M., Caytiles, R.D., and Iyengar, N.C.S.N. (2018). Classification of diabetic retinopathy images by using deep learning models. *International Journal of Grid and Distributed Computing*, **11**(1): 89-106.
[17]. Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, **62**(2): 406-418.
[18]. Basha, S.M., and Rajput, D.S. (2017). Evaluating the impact of feature selection on overall performance of sentiment analysis. In *Proceedings of the 2017 International Conference on Information Technology* (pp. 96-102). ACM.
[19]. Basha, S.M., & Rajput, D.S. (2018). Fitting a Neural Network Classification Model in MATLAB and R for Tweeter Data set. In *Proceedings of International Conference on Recent Advancement on Computer and Communication* (pp. 11-18). Springer, Singapore.

How to cite this article: Basha, S.M., Bagyalakshmi, K., Ramesh, C., Rahim, R., Manikandan, R. and Kumar, A. (2019). Comparative Study on Performance of Document Classification Using Supervised Machine Learning Algorithms: KNIME . *International Journal on Emerging Technologies*, **10**(1): 148-153.