

Comparative Analysis of Selected Text Mining Classifiers

Omokanye S.O.^{1*}, Abikoye O.C., Aro T.O.³, Akande H.B.⁴, Aregbesola K.M.⁵

^{1,2}Department of Computer Science, University of Ilorin, Ilorin, Nigeria

^{3,5}Department of Mathematical and Computing Sciences, KolaDaisi University, Ibadan, Nigeria

¹Department of Telecommunication Science, University of Ilorin, Ilorin, Nigeria

*Corresponding Author: taiwo774@gmail.com Tel.: +2348036176349

Available online at: www.isroset.org

Received: 22/Dec/2020, Accepted: 01/Feb/2021, Online: 28/Feb/2021

Abstract— Text classification is a method of knowledge engineering in which expert level knowledge on classifying documents such that similar documents can be arranged in their respective categories. The field has been a growing area of research as there has been an increase in the availability of textual information. In this paper, three classifiers k-Nearest Neighbour (KNN), Multinomial Naïve Bayes (MNB), and Support Vector Machine (SVM) algorithms were compared based on accuracy and time taken to build a model using five datasets. Experimental results showed that with an increment in the number of training documents, the classification accuracy of the algorithms also increased and algorithms performed better in binary classification tasks than in multiclass classification tasks. The best accuracy of 98.3495% was recorded in SVM using SMS Spam collection dataset compared with other classifiers and datasets, for binary classification, the best classification accuracy of 98.3495% was obtained in SVM using SMS Spam collection dataset, and the best accuracy of 88.28% was obtained in SVM using Reuters 50-50 dataset. The lowest time of 0.01s which was also considered as the best time taken to build a model was recorded in MNB classifier.

Keywords— Naïve Bayes; Nearest Neighbor; Support Vector Machine; Text mining

I. INTRODUCTION

The process that involves the extraction of interesting, intriguing, and nontrivial patterns from a huge amount of text documents is referred to as text mining [1]. The characteristics of documents can be labelled as belonging to a single class or for multiple classes in possible situations. The classification of text is a technique of categorizing a document to a pre-established class [2]. Classification of text is a supervised learning approach in which the algorithms learn from examples to carry out their tasks [3]. It is different from the unsupervised machine learning type in which there is no example from which the algorithm can learn from. If a document can only be given to one class, this is referred to as a single-labelled classification, also if the document can be apportioned to more than one class, this is referred to as a multi-labelled classification [4].

A single-labelled classification of text can be divided into a binary class problem if a data item is given only to one of two classes and becomes a multi-class problem if there are more than two classes in which a data item can be assigned to. Classification of text comprises the representation of a document, selection of feature and transformation of feature, data mining algorithms application and lastly the evaluation of the data mining technique [5].

The classification identifies a technique that predicts the categories of unidentified entities [6], for example, an institution needs to know whether a student in the final year will graduate, the possible two classes are, spill and graduate for the data. The two classes can be denoted using discrete numbers. This classification is referred to as learning based on a supervised approach because the groups have been labelled for the training data.

A machine learning algorithm (MLA) constructs a classifier into two phases [7]. The first phase is the training stage, which produces a model for classification by studying the training data in which the class is labelled and the second phase which is a testing phase that evaluates the learning algorithm accuracy and classification of unidentified entities into respective classes. An MLA creates the model to be employed for classification by considering data in the training stage, in which the classification model is evaluated using data in the testing phase such an evaluation is called accuracy.

LITERATURE REVIEW

Text Classification

Text classification involves the identification of data used in the training of machine learning algorithms, generally, the data formats are not appropriate for mining [8]. This may not produce effective outputs, so these data are pre-processed using various pre-processing methods and at times, manual pre-processing needs to be applied after which learning algorithms are used. Text classification is

being applied in different domains such as in astronomy, business, question analysis, biology, media, journal analysis, commerce, curriculum vitae analysis, medicine, etc. Several MLAs have been used for the classification of text, some are rule-based such as decision trees, and others use statistically-based principles like NB and Bayesian networks. Some algorithms make use of instance-based learning, they learn when the definite classification is conducted. Examples are RepTree, SVM, and Artificial Neural Networks.

K-Nearest Neighbour (KNN)

It is simple to implement and non-parametric classifier, although the classification time of KNN is long, and obtaining the optimal value of k is quite challenging. Among the text classification techniques, KNN is considered to be the best method for text categorization [9]. The Nearest Neighbor rule (NN) is the general type of KNN when $K = 1$. Feature selection and space transformation are very important when k -NN is used because it applies all features in its calculation of distance and thus it should tend to optimize performance. The NN matches a certain test sample to the samples in the training phase related to it [10]. KNN uses the function of distance to decide the members of k that are nearest to the test instance which is not identified.

Support Vector Machine (SVMs)

This is a generalised linear classifier, which is mostly used in classification, regression and prediction tools [11]. SVM applies the concepts of ML to improve the accuracy of the prediction model, which escapes data over-fit. An effective learning approach needs to avoid over-fit of data. SVMs are based on a technique that identifies a distinct type of linear model referred to as maximum margin hyper-plane. The instances that are closest to the maximum-margin hyper-plane, the ones with the minimum distance are called support vectors [12].

Multinomial Naïve Bayes (MNBs)

It uses the Bayes rule and it adopts independence naively [13], it increases events' probabilities considering that they are not dependent. NB has been identified to perform efficiently with some databases and when is being applied with feature selection algorithms, it removes irrelevant and redundant attributes, it works effectively [14], the Naïve Bayes is shown in Equation (1)

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \quad (1)$$

MNB is a model of NB that not only captures the presence or absence of words as in the ordinary NB classifier but also captures the existence of a word in a document [15], MNB is more suitable for text classification as it achieves better performance when the vocabulary size is relatively large as is usually the case of text datasets.

II. RELATED WORK

Many studies have been conducted applying different classification algorithms in text mining:

Naz et al [16] handled the problem of selection of feature subset for classification of sentiments through the ensemble classification. The model combined maximum relevance (mRMR) and the minimum redundancy with Forest Optimization Algorithm (FOA) for feature selection. The output optimization for each classifier was done through the introduction of an ensemble-based method. The evaluation of the model was achieved using datasets in the UCI repository, where the k -NN and NB were the classifiers employed. For the sentiments classification, 15-20% of enhancement was obtained. The classification of sentiments was achieved using Blitzer's dataset consisting of reviews of electronic products. The results were enhanced by an ensemble of k -NN, NB, and the sentiment classification recorded an accuracy of 95% for SVM.

Thangaraj and Sivakami [17] studied different techniques of text classification used in practice. The advantages and weaknesses were analyzed to provide responsiveness concerning the different knowledge extraction possibilities in data mining. Many articles from 2010 to 2017 on the classification of text methods in Artificial Intelligence were chosen from the prominent journals of computer science were analyzed.

Wongso et al [18] developed a news article classification model in the Indonesian language. Data were obtained a web crawling approach from www.cnnindonesia.com. The document was pre-processed by lemmatization and removal of stopwords just to reduce the noise effect in the document. A selection of features was performed by a TF-IDF and SVD algorithm. This was achieved to obtain relevant words inside the document. The classification of selected features was achieved using MNB Multivariate Bernoulli NB, and SVM. The experimental results revealed that the combined MNB and TF-IDF gave outstanding output compared to the other techniques, in which precision value gave 0.9841519 and its recall value gave 0.9840000. The final result obtained outperformed the existing models that recorded an accuracy of 85%.

Bui et al [19] presented the classification of PDF text to control the extraction of information from publication. The study developed an alternate technique for recognition of the PDF structure by the application of PDF Box to obtain raw texts and a multi-pass sieve algorithm (MSA) for classification. The MSA produced better accuracy than commonly used classification algorithms. The MSA also enhanced the IE system performance compared to the extraction of the shelf PDF. Recognition of PDF structure opened the opportunity to research in text mining on PDF files, which is a significant information basis for research in biomedical.

Yang et al [20] came up with a document classification using a hierarchical attention network. The model provided two unique features: a hierarchical structure that showed the documents of hierarchical structure and dual attention mechanisms. This was used at the sentence and

word level to attend differentially to more and less relevant content when constructing the representation of the document. Experiments performed on six large scale classification of text presented that the developed model performed excellently by a considerable margin. The visualization of the attention layers illustrated that the model selected qualitatively informative words and sentences.

Goudjil et al [21] developed an effective method of learning for classification of text. The study selected samples of information to be labelled manually by an expert. SVM was employed to logically choose data that were adequately labelled to minimize the labelling work without effect on accuracy. Results showed that the proposed system considerably decreased the labelling work and enhanced accuracy.

Dey Sarkar et al [23] proposed a dual-phase selection of feature method using a feature selection in univariate form and feature clustering. The model applied the selection of feature technique in univariate type to reduce number of features and then used clustering to choose feature sets that are independent. The comparison of over thirteen datasets and evaluation was achieved to show the approach efficiency. The evaluation of the model revealed that Naive Bayes performed effectively than the remaining classification algorithms. The developed model outperformed conventional algorithms like CFS and greedy search based wrapper.

Nguyen and Shirai [24] built many label classifiers that determined technical paper topics. The model used the knowledge that papers are properly arranged and several portions of papers are more relevant for the classification of text, portion like title, abstract, introduction and conclusion are widely employed in the representation of text. New features; Title Bi-Gram and Title SigNoun were introduced for performance improvement. Experimental results showed that feature selection using segmentation of the text were effective. A new model for classification of text was proposed which used the structure of papers referred to as the Back-off model gave 60.45% Exact Match Ratio and F-measure of 68.75%. It was revealed that the Back-off model performed better than the two existing methods, MLkNN and Binary technique.

III. METHODOLOGY

The comparative analysis of some selected classifiers of text mining involved several processes of text mining as shown in Figure 1. Five datasets: Reuters transcribed subset, Mini 20-Newsgroup, Sentiment labelled instances, Reuters 50-50, and SMS spam collection. Datasets were selected to perform in different sizes and forms of text collections to understand how the algorithms perform with an increase in the size of instances. The data structure used was the Bag of words model and all documents were normalized to ensure equal chances of being retrieved irrespective of document length. TF-IDF was used to

determine important words. The study used K-Nearest Neighbour (KNN), Naïve Bayes Multinomial (MNB) and SVM algorithm for classification. All experiments were conducted in the WEKA data mining tool.

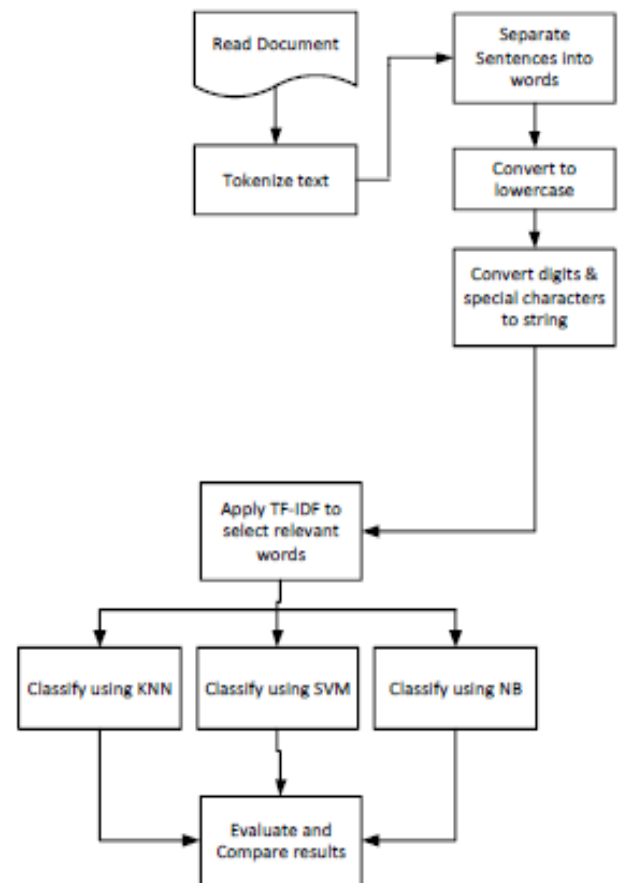


Figure 1: Classification Model Architecture

3.1 Collection of Data

The datasets used in this study are all available in the University of California, Irvine (UCI) database [<http://kdd.ics.uci.edu>].

(i) Reuters transcribed subset: produced by reading out 200files from the 10 largest Reuters classes by native Indian speakers with a Speech Recognition system to create resultant transcriptions. The nature of noise in this dataset is such that due to the pronunciation of the speakers the machine can transcribe the same word differently.

(ii) Mini 20-Newsgroup: Contains 2000 messages partitioned nearly evenly among 20 different newsgroups. Each of the twenty newsgroups has in it 100 UseNet articles for that group.

(iii) Sentiment labelled instances: 3000 instances of texts that have been labelled as either positive or negative sentiment, the texts are extracted from movies, products and restaurant reviews. In collating the dataset, it was ensured that the statement positive or negative so as not to have neutral statements.

(iv) Reuters 50-50: Subsets of the Reuters-21578 collection, fifty authors of texts labelled with at least one subtopic of the class corporate/industrial were obtained. The training corpus contains 2,500 texts and the test corpus includes other 2500 texts (50 per author) non-overlapping with the training texts, but in this experiment, 10-fold cross-validation was used, the combination of training and test sets formed a corpus of 5000 texts.

(v) SMS spam collection: The dataset consists of SMS messages which are classified as either spam or ham. The dataset contains 5574 sentences out of which 4827 are ham messages and 747 are spam messages. The datasets used for the evaluation of the developed classification model vary in different sizes as shown in Table 1.

Table 1: Datasets and Sizes

Dataset	Size in bytes
Reuters transcribed	254kilobytes
Mini 20-Newsgroup	3.63 megabytes
Sentiment labelled instances	207kilobytes
Reuters 50-50	7.55megabytes
SMS spam collection	489kilobytes

3.2 Data Pre-Processing

The datasets were converted to “arff” format for use in WEKA, and in tokenizing the documents the following characters were removed \t,.;:\'()?!@#%^\^ and *()_+|\?><=[]{}[/\~\'". The mini-newsgroup dataset contains emails so the email headers were removed leaving only the content of the mail. It was discovered that the Reuters transcribed subset has 21 text files in its “trade” class instead of the reported 20 per class, so one of the text files was removed at random.

3.3 Feature Selection Phase

This study applied TF-IDF to find the relative word frequency in a document compared to the entire document corpus. Thus it determines the relevance of a specific word in a document.

3.4 Parameters Tuning of Classifiers

KNN: Leave-one-out cross-validation was applied to find the best value of k between 1-100 for the individual datasets as the best k-value depends on the nature and type of data. Ball tree was the NN search algorithm used while Euclidean distance was used as the distance function. SVM: Experiments were conducted on all datasets and the linear kernel was found to give a better result as compared to other kernels available in WEKA on all datasets except the mini-newsgroup dataset in which the radial basis function (RBF) kernel had a slightly higher accuracy but still the same with the linear kernel at a 1% level of statistical significance moreover when a higher number of instances and attributes as in text data, the linear kernel is said to perform best. MNB: MNB has no tuneable parameter.

IV. RESULTS AND DISCUSSION

Experimental analysis was conducted for the classification of text mining using some selected classifiers. The different results were obtained for accuracy and time taken to build the model. The following graphical representations give a comprehensive description of experimental results.

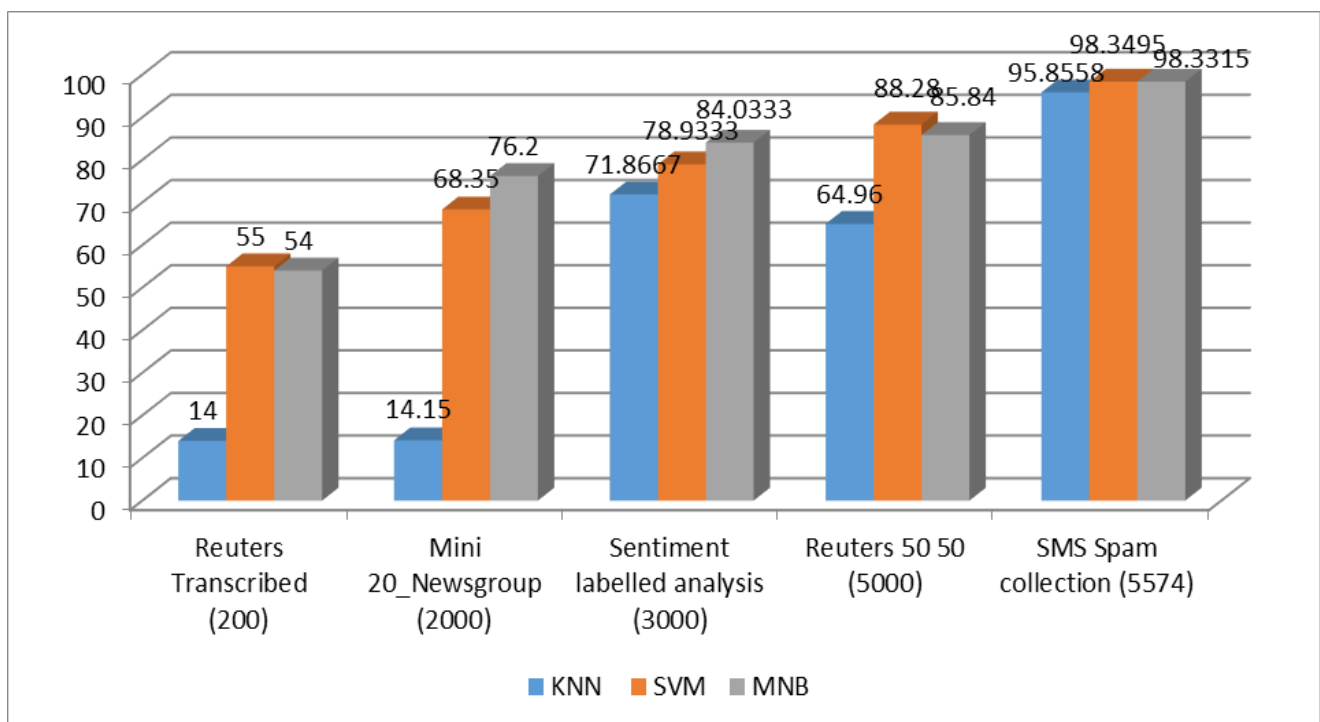


Figure 2: Accuracy on all datasets according to increasing training instances

In Figure 2, the highest classification accuracy of 98.3495% which was recorded as the best accuracy was obtained in SVM using SMS Spam dataset compared with other datasets. Also, the lowest classification accuracy of 14% was recorded in KNN classifier using Reuters Transcribed (200). KNN classifier consistently had the lowest classification accuracy on all datasets. All classification algorithms had the lowest accuracies on the Reuters transcribed dataset and this can be attributed to the noise in the dataset, being a noise caused by transcription incoherence, it reduces the ability of a word to distinguish between the classes when the same word is wrongly transcribed in a similar document. Generally, from the results, it was shown that as the size of the training instances increases, the accuracy of the classifiers also increases.

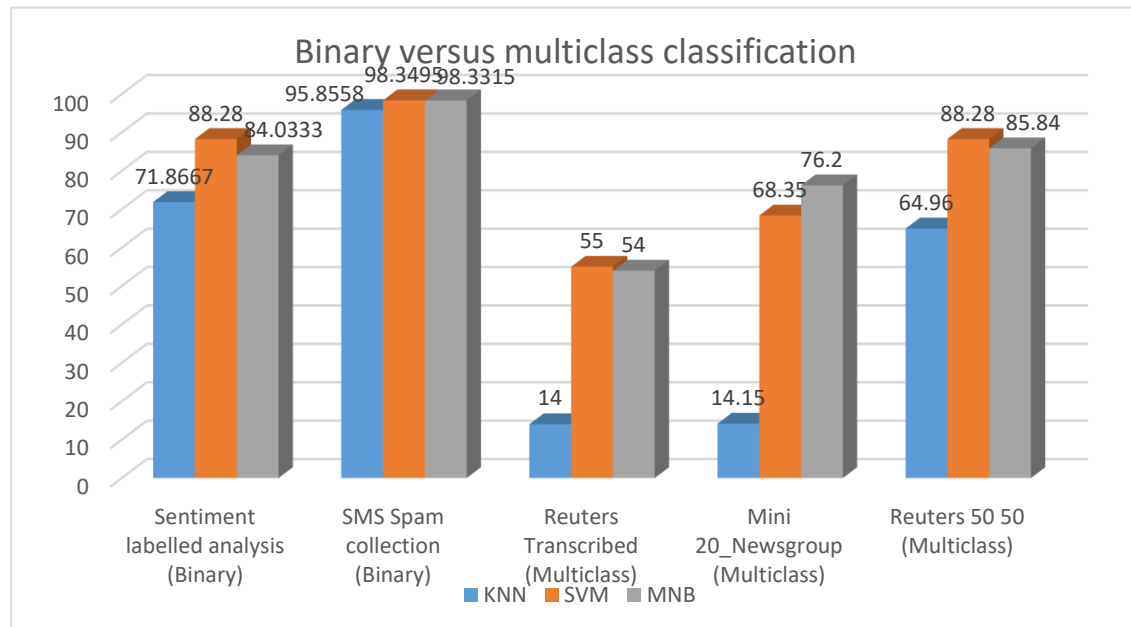


Figure 3: Binary vs Multiclass Classification

From Figure 3, for binary classification, the highest classification accuracy of 98.3495%, which was the best accuracy was obtained in SVM using SMS Spam dataset, and the lowest classification accuracy of 71.8667% was obtained in KNN for Sentiment Labelled Analysis. For the multiclass classification, the highest classification accuracy of 88.28% was recorded in SVM for Reuters 50-50, while the lowest classification accuracy of 14% was obtained in KNN for Reuters Transcribed.

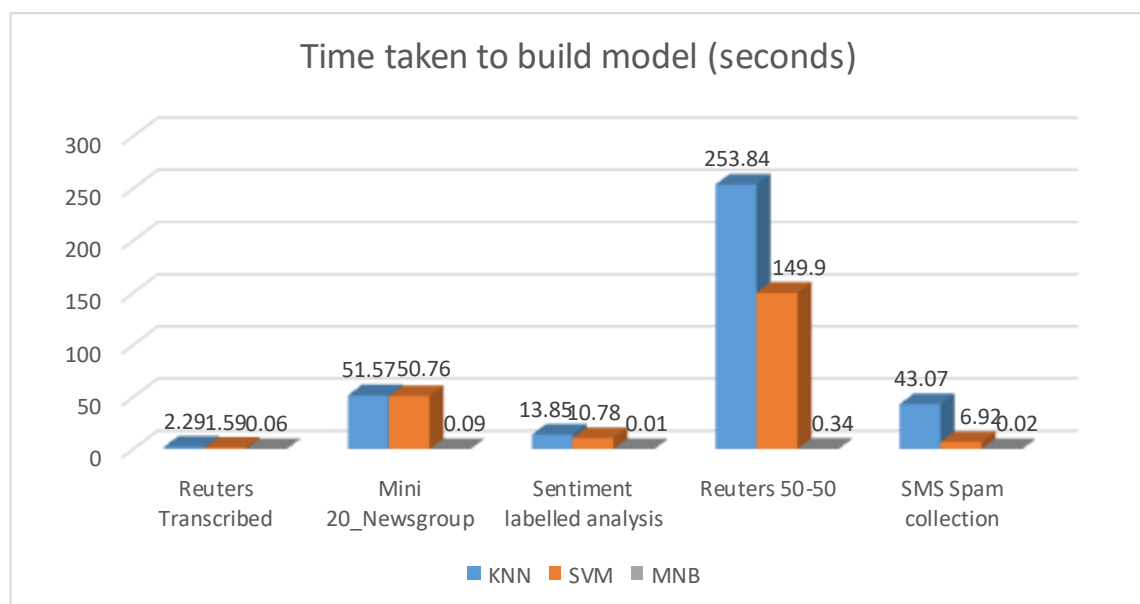


Figure 4: Time taken to Build Classification Model

As shown in Figure 4, the time taken to build the classification model by the different algorithms. The

longest time of 253.84s was taken to build a model using KNN for Reuters 50-50 dataset. Also, the lowest time of

0.01s was recorded in MNB for Sentiment Labelled Analysis. All the classification algorithms took the prolong time in building their model on the Reuters 50-50 dataset followed by the Mini 20-Newsgroup dataset. The reason for this may be found from the manual inspection of the datasets, over 80% of the text instances in each class for the Reuters transcribed, mini 20-Newsgroup and Reuters 50-50 datasets contains more than 100 words each while for the Sentiment Analysis and the SMS spam collection datasets, over 80% of their text instances contains less than 30 words. The main factor responsible for the time taken for an algorithm to build its classification model is the volume of textual data contained in the instances.

V. CONCLUSION

The technique of text Mining involves the selection of relevant patterns in unstructured text to get information that is of high-quality from a text. Undefined or secret information is obtained through the application of text mining process of diverse techniques in mining. This paper conducted a comparative analysis of some selected text mining classifiers. Results showed that the SVM classifier performed effectively in classification of text with the best accuracy of 98.3495% obtained in both binary and multiclass classification using SMS Spam collection and Reuters 50-50 dataset. It was revealed that MNB recorded the lowest time of 0.01s in building its classification model as compared to the other classification algorithms. It was also observed that the availability of more training datasets improved the classification accuracy of algorithms. The amount of textual content in the training documents was found to affect the required time for classification algorithms to build a model.

REFERENCES

- [1] P. Davcheva, "Text Mining Mental Health Forums-Learning from User Experiences," in Twenty-Sixth European Conference on Information Systems, **pp. 1–11, 2018.**
- [2] K. J. Kowsari, K. Meimandi and D. Heidarysafa, M., Mendu, S., Barnes, L., and Brown, "Text Classification Algorithms: A Survey," *Information*, **pp. 1–68, 2019.**
- [3] S. Yogapreethi, N. & Maheeswari, "A Reviewing on Text Mining in Data Mining," *Int. J. Soft Comput.*, **vol.7, no. 2, pp. 1–8, 2016.**
- [4] N. Sailaja, N. V., Padmasree, L., & Mangathayaru, "Survey of Text Mining Techniques, Challenges and their Applications," *Int. J. Comput. Appl.*, **vol. 146, no.11, pp. 30–35, 2016.**
- [5] D. Soni, A., Kumar, V., Kaur, R., & Hemayathi, "Predicting Student Performance Using Data Mining Techniques," *Int. J. Pure Appl. Math.*, **vol. 119, no. 12, pp. 221–227, 2018.**
- [6] R. Shaw and H. Bosworth, "Short message service (SMS) text messaging as an intervention medium for weight loss: A literature review," *Health Informatics J.*, **vol. 18, no. 4, pp. 235–250, 2012.**
- [7] M. Dietterich, T. Bishop, C. Heckerman, D. Jordan, M. and Kearns, "Introduction to Machine Learning Second Edition Adaptive Computation and Machine Learning," Massachusetts Institute of Technology, **2015.**
- [8] A. Kaushik and S. Naithani, "A Comprehensive Study of Text Mining Approach," *Int. J. Comput. Sci. Netw. Security*, **vol. 16, no. 2, pp. 69–76, 2016**
- [9] K. & Nikhath, A. K., Subrahmanyam and R. Vasavi, "Building a K-Nearest Neighbor Classifier for Text Categorization," *Int. J. Comput. Sci. Inf. Technol.*, **vol.7, no. 1, pp.254–256, 2016.**
- [10] M. Moosavian, A.Ahmadi, H. Tabatabaefar, A. and Khazaei, "Comparison of two classifiers; K-nearest neighbour and artificial neural network, for fault diagnosis on the main engine journal-bearing," *Shock Vib.*, **vol. 20, no. 2, pp. 263–272, 2013.**
- [11] K. Matsushita et al., "Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: A collaborative meta-analysis of individual participant data," *Lancet Diabetes Endocrinol.*, **vol. 3, no. 7, pp.514–525, Jul. 2015.**
- [12] A. K. S. Banu and S. H. Ganesh, "A Hybrid Approach for an Efficient Classification Using Decision Tree and SVM," *Int. J. Comput. Sci. Mob. Comput.*, **vol. 7, no. 2, pp. 42–48, 2018.**
- [13] R. Mohana and S. Sumathi, "Document classification using Multinomial Document classification using Multinomial Naïve Bayesian Classifier," *Int. J. Sci. Eng. Technol. Res.*, **vol. 3, no. 5, pp. 1557–1563, 2014.**
- [14] A. Jovic, K. Brkic, and N. Bogunovic, "A Review of Feature Selection Methods with Applications," in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), **pp. 1200–1205. 2015.**
- [15] A. Abbas, M., Kamran, A. M., Abdul, A. J., Memon, S., & Ahmed, "Multinomial Naïve Bayes Classification Model for Sentiment Analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, **vol. 19, no. 3, p. 62, 2019.**
- [16] M. Naz, K. Zafar, and A. Khan, "Ensemble-based classification of sentiments using a forest optimization algorithm," *Data*, **vol. 4, no. 2, pp. 1–13, 2019.**
- [17] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review," *Interdiscip. J. Information, Knowledge, Manag.*, **vol. 13, pp. 117–135, 2018.**
- [18] B. C. and Wongso, R. Luwinda, F. A. Trisnajaya and O. R. Rusli, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, **vol.116, pp. 137–143, 2017.**
- [19] D. D. A. Bui, G. Del Fioli, and S. Jonnalagadda, "PDF text classification to leverage information extraction from publication reports," *J. Biomed. Inform.*, **vol. 61, pp. 141–148, 2016.**
- [20] X. Yang, Z., Yang, D., Dyer, C., He and E. Smola, A., & Hovy, "Hierarchical Attention Networks for Document Classification," 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf., **pp. 1480–1489, 2016.**
- [21] N. Goudjil, M. Koudil, M. Bedda, M. and Ghoggali, "A Novel Active Learning Method Using SVM for Text Classification," *Int. J. Autom. Comput.*, **vol. 15, no. 3, pp. 290–298, 2018.**
- [22] J. Dey Sarkar, S. Goswami, S. Agarwal, A. and Akhtar, "A Novel Feature Selection Technique for Text Classification Using Naïve Bayes," *Int. Sch. Res. Not.*, **pp. 1–10, 2014.**
- [23] A. Dey, "Machine Learning Algorithms: A Review," *Int. J. Comput. Sci. Inf. Technol.*, **vol. 7, no. 3, pp. 1174–1179, 2016.**
- [24] T. H. Nguyen and K. Shirai, "Text classification of technical papers based on text segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **vol. 7934 LNCS. pp. 278–284, 2013.**
- [25] R. Mccue, "A Comparison of the Accuracy of Support Vector Machine and Nave Bayes Algorithms In Spam Classification," Santa Cruz, CA, **2009.**
- [26] P. D. Shahare and R. N. Giri, "Comparative Analysis of Artificial Neural Network and Support Vector Machine Classification for Breast Cancer Detection," **pp. 2114–2119, 2015.**