

Enhanced textual data classification using Particle Swarm Optimization (PSO) algorithm

Taye Oladele Aro

Department of Mathematical & Computing Sciences, Faculty of Applied Sciences
KolaDaisi University, Ibadan, Oyo State
Nigeria
taye.aro@koladaisiuniversity.edu.ng

Hakeem Babalola Akande

Department of Telecommunication, Faculty of Communication and Information Sciences
University of Ilorin, Nigeria
Akande.hb@unilorin.edu.ng

Kayode Sakariyah Adewole

Department of Computer Science, Faculty of Communication and Information Sciences
University of Ilorin, Nigeria
adewole.ks@unilorin.edu.ng

Kehinde Moses Aregbesola

Department of Mathematical & Computing Sciences, Faculty of Applied Sciences
KolaDaisi University, Ibadan, Oyo State
Nigeria
kehinde.aregbsola@koladaisiuniversity.edu.ng

Muhammed Besiru Jibrin

Department of Computer Science, Faculty of Science
Federal University of Kashere, Gombe
Nigeria
bashnjibrin@gmail.com

Abstract

The upturn in digital data acquisition techniques has resulted in a large volume of data. Finding suitable arrangements and trends to analyze the text documents from a huge volume of data remains a serious challenge. The number of irrelevant and redundant features from text data is on the high increase, hence the need to introduce an effective feature selection approach to get the most features that are relevant from the huge text data. This paper applied Particle Swarm Optimisation (PSO) algorithm in selecting important features for accurate text classification. Five classification algorithms: C4.5 Decision Tree, K-Nearest Neighbour (KNN), Multinomial Naïve Bayes (MNB), Rep-Tree (RT), and Radial Basis Function (RBF) were used. The developed text classification model was evaluated using two datasets: SMSSpam and Sentiment Analysis dataset. Experimental results showed that when the PSO was not applied the best accuracy of 98.2239% was obtained in MNB for SMS Spam Dataset, the best accuracy of 84.0333% was obtained in MNB for Sentiment Analysis Dataset. The precision value recorded when PSO was not applied gave the best value of 0.983 in MNB for SMS Spam Dataset and the best precision value of 0.724 in MNB for Sentiment Analysis Dataset. Also, the best recall when PSO was not introduced gave 0.982 in MNB for SMS Spam Dataset, the best recall of 0.840 was obtained in MNB for Sentiment Analysis Dataset. The improvement in results when PSO was applied only showed that the accuracy of 65.2667% in KNN and 71.2333% in RBF for Sentiment Analysis Dataset. The precision recorded an improvement of 0.713 in RBF and also the recalls of 0.653 in KNN and 0.712 in RBF for Sentiment Analysis Dataset. Finally, the study concluded that the MNB as a classifier performed effectively without the application of PSO algorithm in the text data classification in terms of accuracy, precision and recall.

Keywords: Feature Selection, Text Classification, Particle Swarm Optimization, Textual Data

1. Introduction

Text mining is a process which involves the extraction of interesting, intriguing and nontrivial patterns from a huge amount of text documents (Talib et al., 2016). Availability of a large volume of data is one of the predominant problems of data mining (Thangaraj & Sivakami, 2018). During text data analysis, not all the features available are useful for classification purpose (Sulova, Snezhana, Todoranova, Penchev & Nacheva, 2017). Several techniques of feature selection have been proposed such as filter, wrapper and hybrid techniques to reduce high dimensional text data. Studies on text classification are directed towards determining the possibility of representing documents in a more appropriate way, reducing the dimensions of documents, indexing of documents, developing an algorithm to categorize each document depending on criteria in consideration (Zelaia, Alegria, Aregi & Sierra, 2011).

Text classification is a supervised learning technique in which the algorithms learn from examples to carry out its tasks (Yogapreethi & Maheeswari, 2016). It is different from the unsupervised machine learning type in which there is no example from which the algorithm can learn from. There are two dimensions of classification of text task (Sailaja, Padmasree, & Mangathayaru, 2016), one is to classify documents to only a single category (Patel & Soni, 2012), while the other is to classify document into more than one category (Korde & Mahender, 2012). Operations in text classification include representation of documents, feature extraction and/or feature selection, the building of a vector space model, application and evaluation of data mining (Jindal, 2015).

Among several known conventional feature selection techniques, PSO is identified as a nature-inspired evolutionary algorithm for feature selection that does not

converge into a local optimum but global optimal and with a low computational cost (Vashishtha, 2016).

This paper applied the PSO algorithm in the reduction dimension of features through a feature collection of relevant attributes from textual data dataset. After dimensions have been reduced, five classification algorithms were used to categorize both the reduced and the original text data and their results were evaluated and compared.

Related work

Several studies have been carried out using different dimensionality reduction techniques like meta-heuristic optimization algorithms:

Naz, Zafar and Khan, (2019) developed a system to handle the problem of feature subset selection for sentiments classification using ensemble-based classifiers. The study combined the minimum redundancy and maximum relevance (mRMR) with Forest Optimization Algorithm (FOA) of feature selection. The optimization of results for individual classifiers was achieved through the introduction of ensemble-based method. The evaluation of the model was done using UCI repository datasets, where the k-Nearest Neighbor (k-NN) and Naïve Bayes (NB) were the classifiers employed. For the sentiments classification, 15-20% of enhancement was obtained. Also, the Blitzer's dataset consisting of reviews of electronic products was used for classification of sentiments. The results were enhanced by an ensemble of k-NN, NB, and Support Vector Machine (SVM) with an accuracy of 95% for the sentiment classification.

Mowafy, Rezk and El-bakry (2018) presented an effective model for classification of an unstructured text document. The model supported the generality by applying the logical sequence of the process of unstructured text

documents classification step by step and supported the effectiveness by suggesting a well-suited embedded techniques combination for an improved model. The model was evaluated with over 20-Newsgroups dataset in terms of precision, recall and f-score.

The results showed that MNB with TFIDF performed better than KNN as a technique for classification of a text document. The model proposed revealed that the selection of embedded techniques produced robust outputs, enhanced the process of classification and established the compatibility in the techniques selected for various stages.

Al-ab and Al-taani (2017) conducted a comparative analysis by using PSO, which joined the informative scoring with semantic scoring to create a shorter version of an original text in Arabic document with two important evolutionary methods Genetic Algorithm (GA) and Harmony Search (HS). The results showed that PSO achieved a higher precision and F-score measure than the GA and HS techniques.

Devi, Rao, and Setty (2016) developed a system in which PSO and GA were used to select important features in a dengue dataset and Decision Tree (DT) was used as the classifier for the two feature selection algorithms. It was reported that the PSO combined with DT classifier gave a better classification result when compared to GA combined with DT classifier. The result showed a decrease in the error rate of PSO+DT compared to GA+DT.

Muthusamy, Polat, and Yaacob (2015) used two PSO based methods to select features and enhance data to improve the recognition of emotion in speech and glottal signals. PSO based clustering and wrapper based PSO were the two algorithms proposed and when applied to three emotional speech databases, the study reported that PSO was able to classify better when compared with the existing methods.

Saini et al. (2014) came up with a review on the use of PSO and its variants in

addressing the problem of human motion tracking. The study concluded that the algorithm, when used to track pose in a multi-dimensional search space gave better results. Though convergence time is restricted when the search is for global optima, while it was posited that modifying PSO or combining it with some other algorithms produced a better result.

2.1 Classification algorithms

There are several algorithms for classification in text or data mining. These include:

2.1.1 K-NN

It is instance-based learning or lazy learning (Shouman, Turner, & Stocker, 2012) in which function is only locally estimated and all calculation is done during classification (Imandoust & Bolandraftar, 2013). The technique is an essential and simplest classification method in the scenario when there is little or no prior knowledge about the data distribution. The rule holds the complete training set during learning and allocates to each query class denoted by the majority label of its training set (Moosavian et al., 2013).

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. KNN is a non-parametric and simple to implement, 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 (Nikhath, Subrahmanyam & Vasavi, 2016).

2.1.2 Multinomial Naïve Bayes (MNB)

MNB is a variant of Naïve Bayes (NB) that does not only capture the existence or non-existence of words as in the ordinary NB

classifier but also captures the occurrence of a word in a document (Mohana & Sumathi, 2014). The technique is more suitable for text classification as it performs better when the vocabulary size is relatively large as is usually the case of text datasets.

The Naive Bayes approach is a well-recognized method for text classification due to its effective grating assumptions, quick and easy implantation and thus, accuracy can be highly improved with the Multinomial Naive Bayes technique (Abbas, Kamran, Abdul, Memon & Ahmed, 2019).

2.1.3 Decision Tree (DT)

The technique consists of a set of rules which are used sequentially and finally yield a decision using a divide and conquer strategy (Hotho, Nürnberger & Paaß 2005). Generally, the decision tree structure enables the applicability of knowing the component of trained knowledge models. It selects attributes from the root nodes and then creates branches for each possible attribute value (S. Sharma, 2016).

Every branch descending from a node links one of the possible values for this attribute. In DT, classification begins at the root node then goes to the tree branch corresponding to the value of the attribute in the given instance, it reiterates this procedure for the sub-tree rooted at the new node (Yuan, Yuan, Yang, Peng, & Buckles, 2003).

George, Sandhya and Suja (2014) performed a comparative analysis on a wide variety of text classification employing techniques such as DT, Bayesian network and Neural Network for text mining, the study applied classification in text mining in areas like email spam filtering, opinion mining, text filtering of news articles.

2.1.4 Radial Basis Function (RBF)

This is a function that apportions a real value to individual input from its domain (Deepajothi & Selvarajan, 2013), and the value produced by the RBF is always an absolute value. The radial basis functions act

as activation functions. The RBF or RBF kernel, is a common kernel function that have been applied in several kernelized learning methods (Phientrakul & Kijirikul, 2005). In particular, it is commonly used in support vector machine classification.

SVM is one of the robust classification algorithms with an effective generalization ability on unseen data (Kancherla, Bodapati & Veeranjanyulu, 2019). If the data is linearly separable it is possible to draw an infinite number of decision boundaries that separates the data. Sharma and Sharma (2019) developed a text classification method using an ensemble of refined SVM, the merit of the approach is that it can considerably decrease the training data size by using reduction of dimensionality as a pre-training phase.

2.1.5 RepTree (RT)

The technique employs the regression tree logic to produce different iterations of several trees (Mohamed, Salleh & Omar, 2012). It chooses the best one from all trees produced that will be taken as the representative. The RT belongs a fast decision tree learner which constructs a decision/regression tree using information gain as the splitting conditions, and prunes it applying reduced error pruning (Nikhath, Subrahmanyam & Vasavi, 2016).

This technique is based on C4.5 that produces classification of discrete outcome or regression trees of continuous outcome. It builds a regression/decision tree using information gain/variance and prunes it by application of a reduced-error pruning (with back-fitting).

Naji and Ashour (2016) used the RT to enhance text classification, the system tested weighing combinations of schemes to be used in Arabic text data classification tasks.

3. Methods

The developed text classification model consists of different phases, the two publicly

available SMS datasets SMS (Short Messaging Service) Spam Collection and Sentiment analysis dataset were the two databases used in this study. The datasets adequately passed through some pre-processing methods. Particle Swarm Optimization algorithm nature-inspired optimization was introduced to select the most relevant features of text dataset due to its technique of computation in which optimizes a problem by iteratively improving a candidate solution concerning a given measure of quality. Five classification algorithms; C4.5, KNN, RBF, RepTree and MNB were used. The experimental results of the developed text classification model were compared and evaluated while the result of the classification algorithms without data being reduced was used as the baseline method. All experiments were performed using the WEKA 3.9.2 data mining tool.

3.2 Description of datasets

The text data used in this study was obtained from the University of California, Irvine data repository (<http://kdd.ics.uci.edu>). The repository contains a lot of datasets which can be used for research. The first dataset is the SMS (Short Messaging Service) Spam Collection, a set of labelled messages used for investigation of mobile phone spam, having a binary class which is either ham or spam contains 5574 messages with 4827 ham instances and 747 spam instances.

The second dataset is a Sentiment analysis dataset, which consists of product reviews from three websites, amazon, help and imdb websites. The collected dataset combined has 3000 instances with each

website contributing 1000 instances evenly partitioned into 500 positive and five hundred negative reviews respectively. It was ensured that no statement indicating neutrality was added to the dataset.

3.3 Dataset pre-processing

Datasets were converted to “arff” format; this is a required format for the WEKA Data Mining tool to operate on. In tokenizing the documents the following characters were removed `(\r\t,;:\'\\"'")?!\@#\$\%^&*()_+ |\\?><-=[]{}[\^'~\`"-)` and so were not part of the letters making up the words considered by the algorithms in making classification decisions. All documents were normalized and tokenized.

3.3.1 Text normalization

This is a method of converting a text into one standard form. Normalization of text prior storage or processing permits the separation of concerns since the input is guaranteed to be consistent before processes are conducted on it. In this study, stop words were removed using stop words list.

3.3.2 Text tokenization

Tokenization involves the method of splitting a set of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. In the process of tokenization, some characters like punctuation marks are discarded. This study conducted tokenization by converting words to lower cases, separate sentences into words and Remove special characters and conversion of digits into word string.

3.4 Algorithm of PSO

The algorithm for optimization using PSO is shown in Figure 1

Step 1: Start

Step 2: State objective function to be minimized

Step 3: Initialize the parameters c_1 , c_2 , iter max, w, error for PSO

Step 4: Randomly allocate active power to the units satisfying the equality, inequality constraints

Step 5: iteration iter = 0

Step 6: Calculate the objective function

Step 7: Update the gbest and pbest values

Step 8: Update the position and velocity of the particles

Step 9: Iteration iter = iter + 1

Step 10: If the stopping criteria according to errors Go to Step 6 otherwise Go to Step 11

Step 11: gbest of PSO is the solution of economic load dispatch problem

Step 12: End

Figure 1: PSO Algorithm

3.5 Metrics for performance evaluation

The metrics for the evaluation of textual data classification are briefly discussed in the sub-sections.

$$\text{Recall} = \frac{TP}{TP+FP}$$

(2)

3.5.1 Accuracy

It is the percentage of correctly classified instances. It is one of the metrics for evaluating the classification models as shown in Equation (1)

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%$$

(1)

3.5.3 Precision

This is a measure of the proportion of actual negatives that are correctly classified, it is also referred to as positive predictive value as shown in Equation (3)

$$\text{Precision} = \frac{TP}{TP+FP}$$

(3)

3.5.2 Recall

It measures the proportion of actual positives that are correctly classified, this is also referred to as sensitivity which is the true positive rate as given in Equation (2)

Where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative

4. Results and discussion

4.1 Experimental results

This section discusses the results obtained from the two datasets: SMS Spam and Sentiment Analysis Dataset. The comprehensive results in terms of accuracy, precision and recall for both original and reduced datasets with PSO are presented in subsequent sections:

4.1.1 Results of developed text classification model in term of accuracy

Different results were obtained for accuracy with original and reduced datasets with PSO as feature selection based on the results of the five classification algorithms employed as shown in Table 1 and Table 2.

Table 1 shows the result of text classification when no feature selection technique was applied. The highest accuracy of 98.2239% was obtained in MNB for SMS

Spam dataset and 84.0333% in MNB for Sentiment Analysis dataset.

From Table 2, when the particle swarm optimization algorithm was employed on two datasets, the highest accuracy of 96.5016% was obtained in C4.5 DT for SMS Spam dataset and 71.2333% in RBF for Sentiment Analysis.

4.1.2 Results of developed text classification model in term of precision

The results of precision for no feature selection and with PSO as feature selection using the two text datasets; SMS Spam and Sentiment Analysis dataset based on classification algorithms are presented in Table 3 and Table 4 respectively.

From Table 3, the best precision of 0.983 was achieved in MNB for SMS Spam dataset and 0.841 was obtained in RepTree for Sentiment Analysis when no feature selection was used on the two datasets.

Table 1: Accuracy of Developed Text Classification Model (No Feature Selection)

Classifier	Dataset	
	SMS Spam (%)	Sentiment Analysis (%)
MNB	98.2239	84.0333
RepTree	96.5913	72.2667
KNN	95.9993	62.6333
RBF	92.1062	56.8333
C4.5	96.8066	68.5667

Table 2: Accuracy of Developed Text Classification (PSO Feature Selection)

Classifier	Dataset	
	SMS Spam (%)	Sentiment Analysis (%)
MNB	95.7481	60.2333
RepTree	96.0352	50.9000
KNN	95.9993	65.2667
RBF	91.4245	71.2333
C4.5	96.5016	67.7667

Table 3: Precision of Developed Text Classification (No Feature Selection)

Classifier	Dataset	
	SMS Spam	Sentiment Analysis
MNB	0.983	0.724
RepTree	0.965	0.841
KNN	0.960	0.697
RBF	0.922	0.578
C4.5	0.968	0.690

Table 4: Precision of Developed Text Classification Model (PSO Feature Selection)

Classifier	Dataset	
	SMS Spam	Sentiment Analysis
MNB	0.960	0.641
RepTree	0.959	0.600
KNN	0.962	0.653
RBF	0.909	0.713
C4.5	0.964	0.683

From Table 4, the best precision value of 0.964 was achieved in C4.5 DT for SMS Spam dataset and 0.713 was obtained in RBF for Sentiment Analysis when particle swarm optimization algorithm was introduced to select relevant attributes for classification.

4.1.3 Results of Developed Text Classification Model in Term of Recall

The results of recall for no feature selection and with PSO as feature selection using two text datasets; SMS Spam and Sentiment Analysis dataset based on classification algorithms as shown in Table 5 and Table 6.

From Table 5, the best recall value of 0.982 was achieved in MNB for SMS Spam dataset and 0.840 was obtained in MNB for Sentiment Analysis when no feature selection was introduced to select relevant attributes before classification.

In Table 6, the best recall value of 0.965 was achieved in C4.5 DT for SMS Spam dataset and 0.712 was obtained in RBF for Sentiment Analysis when particle swarm optimization feature selection was introduced to select relevant attributes before classification.

Table 5: Recall of Developed Text Classification Model (No Feature Selection)

Classifier	Dataset	
	SMS Spam	Sentiment Analysis
MNB	0.982	0.840
RepTree	0.966	0.723
KNN	0.960	0.626
RBF	0.921	0.568
C4.5	0.968	0.686

Table 6: Recall of Developed Text Classification Model (PSO Feature Selection)

Classifier	Dataset	
	SMS Spam	Sentiment Analysis
MNB	0.957	0.602
RepTree	0.960	0.599
KNN	0.960	0.653
RBF	0.914	0.712
C4.5	0.965	0.678

4.2 Results of Comparative Analysis of the Developed Text Data Classification Model

The section conducts results for comparative analysis on the influence of PSO on text

classification using two text datasets: SMS Spam Dataset and Sentiment Analysis Dataset. The comparison system was based on accuracy, precision and recall as illustrated in the following subsections.

4.2.1 Result of Comparative Analysis of Text Classification Model in Term of Accuracy

This phase shows the results of comparative

analysis of PSO effect on the accuracy of text classification using two text datasets: SMS Spam Dataset and Sentiment Analysis Dataset as shown in Figure 1 and Figure 2.

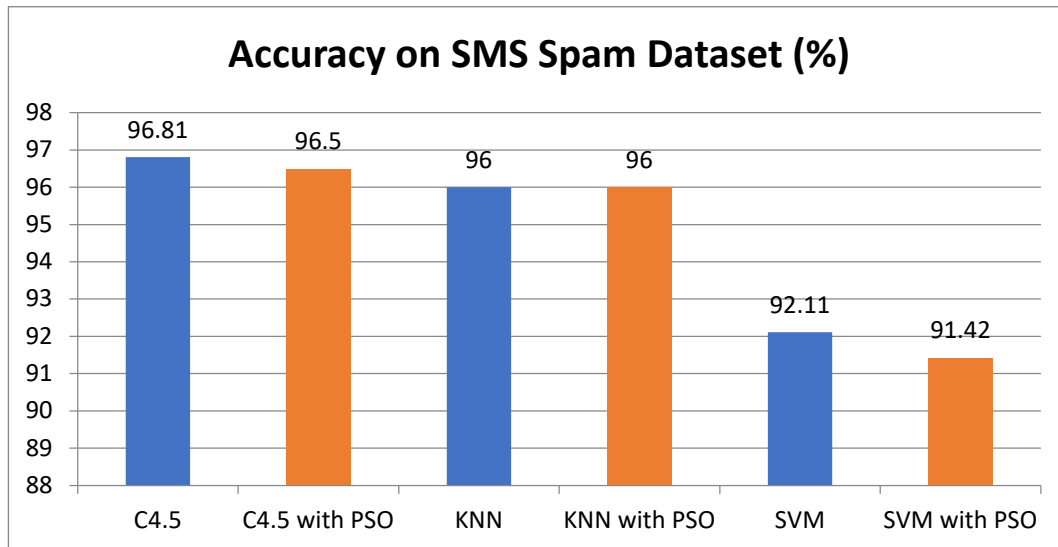


Figure 2: Accuracy on SMS Spam Dataset

In Figure 2, no effect of particle swarm optimization algorithm was recorded on the text data classification accuracy for all the machine learning algorithms with SMS

Spam Dataset. It was shown that the accuracy of text data classification performed effectively well when PSO was not used to reduce the attributes of the dataset.

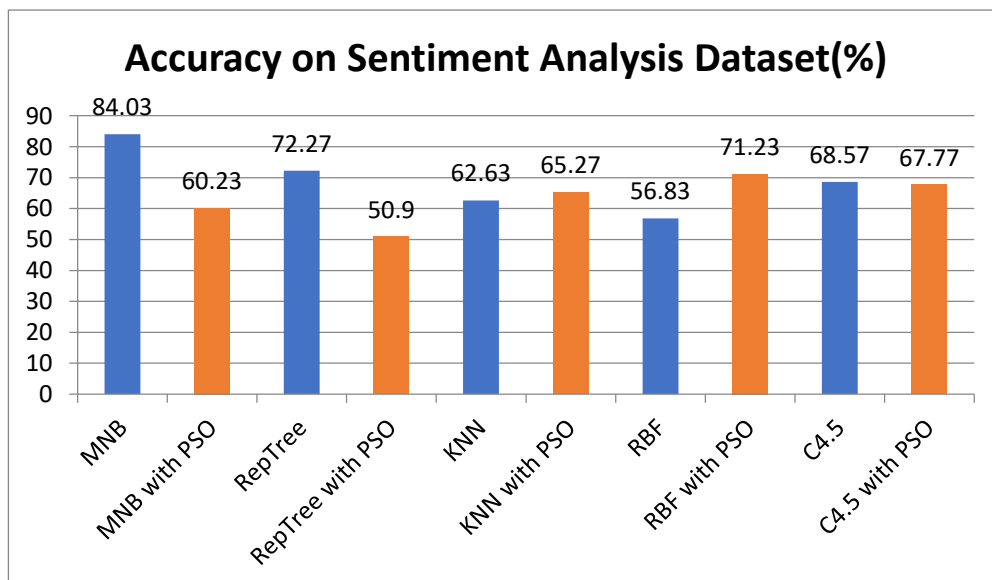


Figure 3: Accuracy on Sentiment Analysis Dataset

From Figure 3, using Sentiment Analysis Dataset, the effect of particle swarm optimization algorithm was only observed in KNN classifier with 65.27% accuracy as compared with 62.63% when PSO was not applied to select relevant attributes. Also, the impact of PSO was felt in RBF with 71.23% accuracy obtained as against 56.83% when PSO was not used.

4.2.2 Result of Comparative Analysis of Text Classification Model in Term of Precision

The section shows the comparative analysis results of PSO effect on the precision of text classification using two text datasets: SMS Spam Dataset and Sentiment Analysis Dataset as shown in Figure 4 and Figure 5.

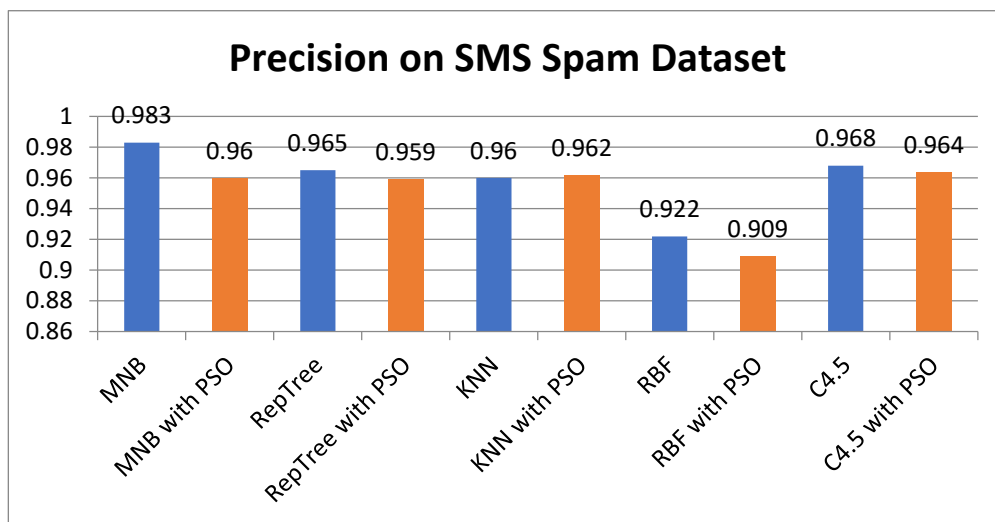


Figure 4: Precision on SMS Spam Dataset

In Figure 4, evaluating with SMS Spam Dataset, the application of particle swarm optimization algorithm had a slight effect on the precision value of KNN Classifier.

While no effect was recorded in precision values of other four machine learning algorithms

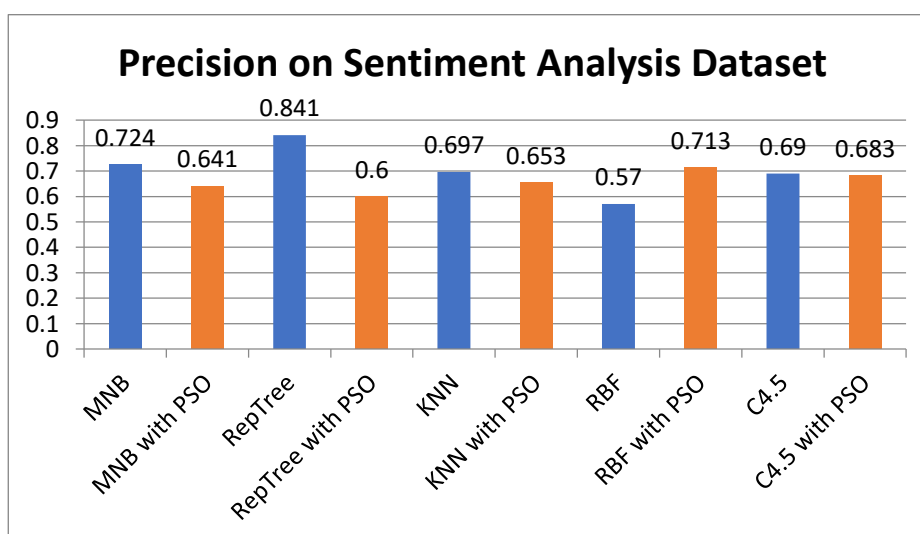


Figure 5: Precision on SMS Spam Dataset

From Figure 5, evaluating with Sentiment Analysis Dataset, the application of particle swarm optimization algorithm had a slight effect only in RBF classifier with the precision value of 0.713 compared to 0.570 when no PSO was applied using the same Classifier. Other classifiers: MNB, RepTree, KNN and C4.5 did not show any effect. From Figure 6, evaluating with SMS Spam Dataset, the application of particle swarm optimization algorithm had no effect on

recall value for all the machine learning algorithms used for classification.

In Figure 7, evaluating the text classification model with Sentiment Analysis Dataset, the application of particle swarm optimization algorithm gave a slight effect only in KNN classifier with recall value of 0.653 compared to 0.626 when no PSO was applied. Other classifiers: MNB, RepTree, RBF and C4.5 did not show any effect.

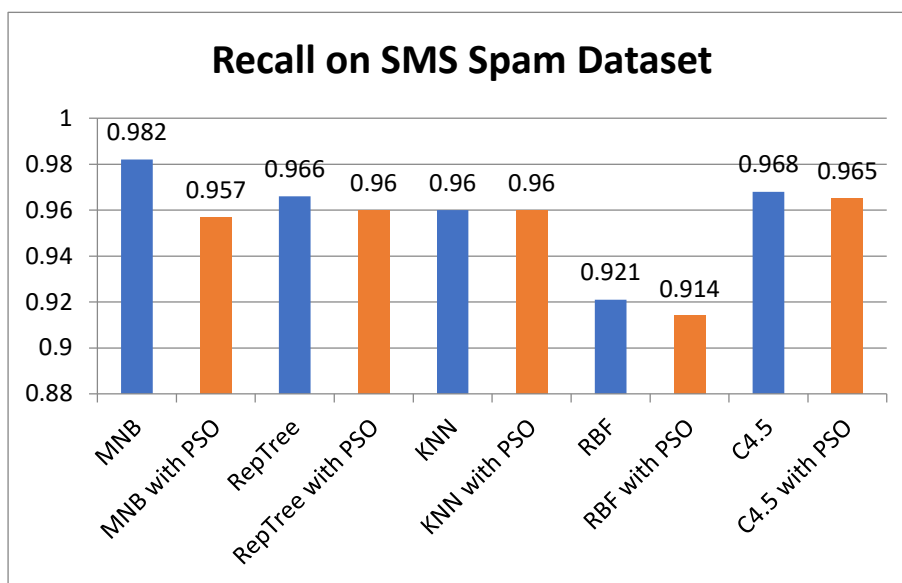


Figure 6: Recall on SMS Spam Dataset

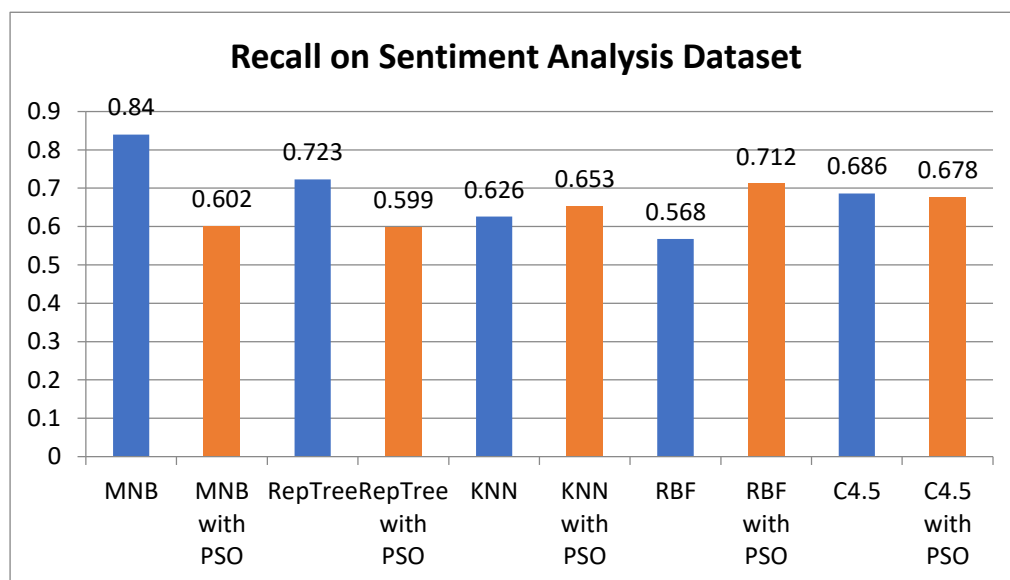


Figure 7: Recall on Sentiment Analysis Dataset

5. Conclusion

This paper investigated the effect of the PSO algorithm in the selection of the most relevant and discriminant features from text datasets before the text classification process. The selected features were passed into five classification algorithms; C4.5, KNN, RepTree and RBF. Results revealed the effect of PSO on the two text datasets: SMS Spam and Sentiment Analysis Dataset. The effect of PSO on accuracy was only showed in KNN with 65.2667% and 71.2333% in RBF for only Sentiment Analysis Dataset, while the impact of PSO for precision value was recorded in RBF with the value of 0.713 for Sentiment Analysis Dataset and the effect of PSO on recall value was obtained in KNN with recall value of 0.653. Overall experimental results showed that PSO as an effective nature-inspired optimization algorithm did not play a significant role in the performance of text classification.

References

- Abbas, M., Kamran, A. M., Abdul, A. J., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. *IJCSNS International Journal of Computer Science and Network Security*, 19(3), 62. <https://doi.org/10.13140/RG.2.2.30021.40169>
- Al-ab, R. Z., & Al-taani, A. T. (2017). ScienceDirect ScienceDirect 3rd International Conference on Arabi Dubai, the United Arab Emirates Arabic Single-Do Document Text Summarization on Using Particle icle Sw Swarm Optimization Algorithm. *Procedia Computer Science*, 117, 30–37. <https://doi.org/10.1016/j.procs.2017.10.091>
- Deepajothi, S., & Selvarajan, S. (2013). Performance Evaluation of SVM--RBF Kernel for Classifying ECoG Motor Imagery. *International Journal of Computer Science and Telecommunications*, 4(5), 44–48.
- Devi, B. R., Rao, K. N., & Setty, S. P. (2016). Towards Better classification using Improved Particle Swarm Optimization and Decision tree for Dengue Datasets. *International Journal of Soft Computing*, 11(1), 18–25.
- George, J., Sandhya, N., & Suja, G. (2014). Classification problem in text mining. *International Journal of Innovative Research in Advanced Engineering*, 1(8), 333–341.
- Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20, 19–62. <https://doi.org/10.1111/j.1365-2621.1978.tb09773.x>
- Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background. *Int. Journal of Engineering Research and Applications*, 3(5), 605–610.
- Jindal, R. (2015). Techniques for text classification : Literature review and current trends. *Webology*, 12(2), 1–28.
- Kancherla, D., Bodapati, J. D., & Veeranjanyulu, N. (2019). Effect of different kernels on the performance of an SVM based classification. *International Journal of Recent Technology and Engineering*, 7(5), 1–6.
- Korde, V., & Mahender, C. N. (2012). Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications*, 3(2), 85–99. <https://doi.org/10.5121/ijaia.2012.3208>
- Mohamed, W. W., Salleh, M. N., & Omar, A. H. (2012). A comparative study of Reduced Error Pruning method in decision tree algorithms. *Proceedings - 2012 IEEE International*

- Conference on Control System, Computing and Engineering, ICCSCE 2012*, 392–397. <https://doi.org/10.1109/ICCSCE.2012.6487177>
- Mohana, R., & Sumathi, S. (2014). Document classification using Multinomial Naïve Bayesian Classifier. *International Journal of Science, Engineering and Technology Research*, 3(5), 1557–1563.
- Moosavian, A., Ahmadi, H., Tabatabaeefar, A., & Khazaei, M. (2013). Comparison of two classifiers; K-nearest neighbour and artificial neural network, for fault diagnosis on the main engine journal-bearing. *Shock and Vibration*, 20(2), 263–272. <https://doi.org/10.3233/SAV-2012-00742>
- Mowafy, M., Rezk, A., & El-bakry, H. M. (2018). An Efficient Classification Model for Unstructured Text Document. *American Journal of Computer Science and Information Technology*, 06(01), 1–9. <https://doi.org/10.21767/2349-3917.100016>
- Muthusamy, H., Polat, K., & Yaacob, S. (2015). Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals. 1–20. <https://doi.org/10.1371/journal.pone.0120344>
- Naji, H., & Ashour, W. (2016). Text Classification for Arabic Words Using Rep-Tree. *International Journal of Computer Science and Information Technology*, 8(2), 101–108. <https://doi.org/10.5121/ijcsit.2016.8208>
- Naz, M., Zafar, K., & Khan, A. (2019). Ensemble-based classification of sentiments using a forest optimization algorithm. *Data*, 4(2), 1–13. <https://doi.org/10.3390/data4020076>
- Nikhath, A. K., Subrahmanyam, K., & Vasavi, R. (2016). Building a K-Nearest Neighbor Classifier for Text Categorization. *International Journal of Computer Science and Information Technologies*, 7(1), 254–256.
- Patel, F. N., & Soni, N. R. (2012). Text mining: A Brief survey. *International Journal of Advanced Computer Research*, 2(4), 2–7.
- Phientrakul, T., & Kijssirikul, B. (2005). Evolutionary strategies for multi-scale radial basis function kernels in support vector machines. *GECCO 2005 - Genetic and Evolutionary Computation Conference*, 905–911. <https://doi.org/10.1145/1068009.1068160>
- Sailaja, N. V., Padmasree, L., & Mangathayaru, N. (2016). Survey of Text Mining Techniques, Challenges and their Applications. *International Journal of Computer Applications*, 146(11), 30–35.
- Saini, S., Rohaya, D., Awang, B., Zakaria, M. N. B., & Sulaiman, S. B. (2014). A Review on Particle Swarm Optimization Algorithm and Its Variants to Human Motion Tracking. 2014, 13–15.
- Sharma, S. (2016). Review on Text Mining Algorithms. 134(8), 39–43.
- Sharma, S. K., & Sharma, N. K. (2019). Text classification using an ensemble of non-linear support vector machines. *International Journal of Innovative Technology and Exploring Engineering*, 8(10), 3169–3174. <https://doi.org/10.35940/ijitee.J9520.0881019>
- Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. *International Journal of Information and Education Technology*, 2(3), 220–223. <https://doi.org/10.7763/IJiet.2012.V2.114>
- Sulova, Snezhana., Todoranova, L., Penchev, B., & Nacheva, R. (2017). Using text mining to classify research papers. *International Multidisciplinary*

Taye Oladele Aro, Hakeem Babalola Akande, Kayode Sakariyah Adewole, Kehinde Moses Aregbesola and Muhammed Besiru Jibrin: Enhanced textual data classification using Particle Swarm Optimization (PSO) algorithm

Scientific GeoConference, 1–9.
<https://doi.org/10.5593/SGEM2017/21/S07.083>

- Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *IJACSA) International Journal of Advanced Computer Science and Applications*, 7(11), 414–418. www.ijacsa.thesai.org
- Thangaraj, M., & Sivakami, M. (2018). Text classification techniques: A literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, 13, 117–135. <https://doi.org/10.28945/4066>
- Vashishtha, J. (2016). *Particle Swarm Optimization based Feature Selection*. 146(6), 11–17.
- Yogapreethi, N. & Maheeswari, S. (2016). A Reviewing on Text Mining in Data Mining. *International Journal on Soft Computing*, 7(2), 1–8. <https://doi.org/10.5121/ijsc.2016.7301>
- Yuan, X., Yuan, X., Yang, F., Peng, J., & Buckles, B. P. (2003). *Gene Expression Classification: Decision Trees vs. SVMs*. 92–96.
- Zelaia, A., Alegria, I., Arregi, O., & Sierra, B. (2011). A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. *Applied Soft Computing Journal*, 1–10. <https://doi.org/10.1016/j.asoc.2011.06.002>.

Authors' Brief

Taye Oladele Aro

He obtained B.Sc. in Computer Science, M.Sc. in Mathematics/Computer Science Option and a Ph.D in Computer Science from the University of Ilorin. He is an academic staff in the

Department of Mathematical and Computing Sciences, KolaDaisi University, Ibadan, Oyo State. His research areas are Data Science, Biometrics, Bioinformatics and Machine learning.

Hakeem Babalola Akande

Hakeem Babalola Akande received B.Sc and M.Sc. in Computer Science, University of Ilorin. He is an academic staff in the Department of Telecommunication Science, University of Ilorin. Research interests include Data Science, Networking, Network Security and Internet of Things.

Kayode Sakariyah Adewole

He received B.Sc and M.Sc in Computer Science from University of Ilorin and Ph.D. from University of Malaya, Kuala Lumpur, Malaysia. He is an academic staff in the Department of Computer Science, University of Ilorin. His research interests include Network Security, Biometrics, Machine learning, Data Streaming, Artificial Intelligence, Big Data Analytics and Internet of Things.

Kehinde Moses Aregbesola

Kehinde Moses Aregbesola obtained B.Sc and M.Sc in Computer Science from University of Ibadan and Ph.D. from Ladoke Akintola University Ogbomoso, Oyo State. He is a Senior lecturer and Acting Head of the Department of Mathematical and Computing Sciences, KolaDaisi University, Ibadan, Oyo State. His research interests are Software Engineering, Data Science and Artificial Intelligence.

Muhammed Besiru Jibrin

He received B.Sc. in Computer Science from Kogi State University, Anyigba, Kogi State and M.Sc. in Computer Science from University of Ilorin, Ilorin. He is an academic staff of Federal University of Kashere, Gombe State. His research domains are Data Science and Artificial Intelligence