AJBAR Vol. 3(4), August 2024: 87-95, ISSN: 2811-2881

Arid Zone Journal of Basic and Applied Research

Faculty of Science, Borno State University
Maiduguri, Nigeria

Journal homepage: https://www.azjournalbar.com

Research Article

Ensemble Machine Learning Algorithm for Telegram Spam Detection

Abubakar Hassan[1*], Muhammad Abatcha[1] and Emmanuel Gbenga Dada[2]
[1] Department of Computer Engineering, University of Maiduguri, Maiduguri, Borno State.
[2]Department of Computer Science, University of Maiduguri, Maiduguri, Borno State
*Corresponding author's Email: abubakarhassan@unimaid.edu.ng, doi.org/10.55639/607.06050402

ARTICLE INFO:

Keywords:
Ensemble Learning,
Machine Learning,
Random Forests,
Logistic
Regression,
Telegram

ABSTRACT

Telegram is one of the popular Instant Messaging (IM) platforms widely used today in the world. It is largely used due to its advantages of high speed, strong security and good privacy, key features for both public and private messaging. But telegram spam messages have become a significant concern for users which not only inundates user's storage space but also exposes them to security threats, including malicious content and phishing attacks. This leads to the need to develop more effective spam detectors for the modern digital technology platforms. Machine learning algorithms have proved to be a robust approach for solving the problem of spam messages causing concerns to internet users. This paper proposed an ensemble machine learning algorithm for Telegram Spam Detection using Random Forests and Logistic Regression as base learners. Experiments were conducted in jupyter environment (Python 3) using dataset obtained from Kaggle. The models were evaluated using the metrics: accuracy, precision, recall, and the F1 measure, which offer a holistic appraisal of the model's efficacy. Experimental results demonstrated that the proposed ensemble model and the Random Forests algorithm achieved 94% accuracy compared to the Logistic Regression model (93%) on the benchmark dataset.

Corresponding author: Abubakar Hassan, **Email:** abubakarhassan@unimaid.edu.ng
Department of Computer Engineering, University of Maiduguri, P.M.B. 1069, Maiduguri, Nigeria

INTRODUCTION

In recent years, spam has become an increasingly great concern on the internet. Spam stops good use of time, storage capacity and network bandwidth by users. It led to unprecedented financial loss a great number of users who were scammed to provide sensitive information such as Personal Identification Number (PIN), Bank Verification Number (BVN) and credit card number (Dada et al., 2019). This is because given the fact that the Internet connectivity and digital technology has become more and more accessible over the world in the last decades, going from 20% of the world population with Internet access in 2005 to 63% in 2021.This statistical data shows that about 4.9 billion people connected to the internet (Castano *et al.,* 2023).

To manage effectively the concerns created by spam, leading digital technology companies in the world such as Google and Microsoft have used machine learning (ML) algorithms for spam filtering. These ML approaches have the ability to learn and classify spam messages by analyzing huge datasets. Furthermore, several ML methods have been proposed by the academic and research community to detect spam messages in emails and social networking platforms ((Dada *et al.,* 2019). Authors in (Dada & Joseph, 2018) proposed a random forest algorithm for email spam classification. It achieved a classification accuracy of 99.92% on the benchmark dataset. In another similar study, (Dada & Bassi, 2018) developed an email spam filter using Logistic Model Tree (LMT) Induction. The technique achieved a classification accuracy of 99.305% on the benchmark dataset. Naïve Bayes and K-Nearest classifiers were implemented by (Pinandito *et al.,* 2017) to detect spam in Twitter trending topics. The Naïve Bayes and K-Nearest Neighbour algorithms detected spam and ham contents with 82% and 71% accuracy respectively.

A hybrid Twitter spam detection technique was proposed by (Kumar et al., 2022). it was found that the random forest algorithm with accuracy (99.26), recall (99.07) and precision (99.49) performed better compared to other algorithms while the Naïve Bayes with accuracy (59.92), recall (98.13) and precision (56.47) performed the least.

In (Baaqeel & Zagrouba, 2020), different supervised ML classifiers were employed to detect SMS spam messages. Support vector Machine (SVM) achieved the highest precision and K-Nearest Neighbour (KNN) has the least performance compared to other classifiers. Author in (Oh, 2021) Proposed ensemble learning model for spam detection in YouTube. Experiments were carried out with six different machine learning techniques using data from popular music videos. In (Mambina *et al.,* 2024), the efficacy of deep-learning models for filtering Swahili SMS spam based on linguistics and behavioral patterns using a real-world dataset from telecommunications companies in Tanzania was investigated. The models were trained and tested with 10 k-fold cross-validation.

The experimental results show that the CNN-LSTM-LSTM hybrid model attained the highest accuracy of 99.98 on the Swahili dataset while CNN-BiLSTM performed better on the UCI dataset with an accuracy of 98.38. Virtually all the ML algorithms for spam detection proposed in the literature were based on email, SMS, twitter and YouTube messages. None have considered detecting spam messages in Telegram. Telegram is a popular instant messaging platform that started in 2013, with more than half a billion active users by 2021(Morgia *et al.,* 2021). It is widely adopted because of its high speed, strong security and good privacy, key features for both public and private messaging (Dargahi Nobari *et al.,* 2021). Hence, this study aims to propose an ensemble machine learning algorithm for spam detection in Telegram.

METHODOLOGY

Dataset Description

The Telegram spam dataset used in this study was obtained from Kaggle. The dataset contains 20,000 messages which can be classified into spam or ham (70-30%).

Data Preprocessing

Pre-processing clean and normalise the text data to ensure that it is in a consistent format which helps enhance model performance. This involves steps such as lowercasing, removing special characters and punctuation, removing stop words, tokenization and stemming. Normalization is a technique employed to homogenize the measure of autonomous

variables or attributes of data. It is usually carried out in the data pre-processing phase.

Feature Extraction

Feature extraction converts the pre-processed text data into numerical features that can be used by machine learning algorithms. It can also be seen as the process of selecting a subset of the terms in the training set and exploiting only this subset as features in text classification. This is accomplished by using some set of rules. Feature extraction makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary and usually enhances classification accuracy by removing noise features. Some of the important telegram features we used for our spam detecting include Message body and subject, Volume of the message, Occurrence count of words, Number of semantic discrepancies patterns in the message, Bag of words from the message content, more blank lines in body.

Machine Learning Algorithms

Machine learning is a subfield of Artificial Intelligence (AI) which is deeply rooted in Mathematics, Statistics and Computer Science (Dada et al., 2022; Mienye & Sun, 2022).With advances in computing, ML algorithms have become increasingly robust and accurate in making predictions and smart decisions. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed and is generally referred to as the most popular latest technologies in the fourth industrial revolution (Domor Mienye & Jere, 2024; Oyewola & Dada, 2022; Sarker, 2021). It is the automatic version of the knowledge forming (Ahmed et al., 2023). Random Forests (RF) and Logistic Regression (LR) machine learning algorithms were employed in this study. RF has a shorter training time and a higher classification accuracy than many of the popular machine learning techniques. The choice of Logistic Regression algorithm is rooted in its simplicity, ease of interpretation, and appropriateness for binary classification.

Random Forests (RF)

Random Forest was developed by Breiman and Cutler. The RF algorithm is an ensemble of decision trees. RF is a meta estimator that fits many classifying decision trees on several sub-samples of the dataset and employs averaging to enhance its predictive accuracy and regulate over-fitting. It employs the bagging technique to build multiple decisions trees using strapped samples. It is a supervised learning algorithm. It is simple, diversified, and can be implemented easily. It does not need a considerable number of resources like time, processing power, or memory before producing optimal solutions to any problem (Mienye & Jere, 2024; Dada *et al.,* 2021; Mienye & Sun, 2022).

RF has proven to perform excellently in solving several real-world problems. It is a good example of ensemble machine learning and regression method suitable for finding solutions to classification and prediction jobs. RF can produce optimal results most of the time, even without tuning any value of the used parameter during the learning process. The benefits of using Random forests include minimized prediction error and improved f-scores compared to several other machines learning algorithms. Besides, its overall performance is better than that of Naïve Bayes and SVMs. Unlike SVM and Neural Networks, RF has a shorter training time. RF has a higher classification accuracy than many of the popular machine learning techniques. The random forest algorithm follows the parallel ensemble learning architecture where the base learners are decision trees (Dada *et al.*, 2021; Mienye & Sun, 2022).

Logistic Regression (LR)

The choice of Logistic Regression algorithm is rooted in its simplicity, ease of interpretation, and appropriateness for binary classification. The aim of logistic regression is to find the best fitting that describe the relationship between dichotomous features. In other algorithm, the goal is to select parameters that minimize the sum of squared errors like in Naïve Bayes. However, logistic regression chooses parameters that maximize the likelihood of observing the sample values (Baaqeel & Zagrouba, 2020).

Proposed Model

Ensemble learning is a machine learning method that combines predictions from two or more base models. Ensemble learning models train two or more base learners and combine their predictions to achieve enhanced performance and greater generalization capacity than the individual base learners. The key motivation behind ensemble learning is the fact that machine learning algorithms have shortcomings and can make errors. As a result, ensemble learning aims to enhance

classification result by utilizing the strengths of numerous base models. Ensemble learning methods are broadly grouped into boosting, bagging, and stacking (Dada et al., 2024; Mienye & Sun, 2022).

Stacking (Stacked generalization) is an ensemble learning strategy that trains individual ML algorithms to combine the predictions from multiple ensemble members. It was introduced by Wolpert in 1992 to reduce the generalization error in machine learning problems. Stacking is useful in situations where many ML algorithms are uniquely great on a specific problem. Then the stacking framework would employ a separate ML model to learn when to use the predictions from the different models. Specifically, it involves building models using different base learners (level-0 models) and a meta-learning algorithm that trains another model to combine the predictions from the base algorithms. Meta-learning is the subset of machine learning where algorithms are trained using output of other ML models and make more accurate predictions given the predictions made by the other base algorithms (Dada *et al.*, 2024; Mienye & Sun, 2022).

This study employed stack ensemble learning (SEL) algorithm to detect spam messages in Telegram. The stacking algorithm architecture incorporates two base learner models: Random Forests and Logistic Regression. In addition, it includes a majority voting meta-learner that combines the predictions from the base algorithms. The level 0 models serve as the foundational models, while the level 1 model functions as the meta-model. The stacking ensemble algorithm comprises the initial training data, primary level models, primary level predictions, secondary level models, and the final prediction. The underlying framework of the stacking architecture is as follows:

- Original data: The dataset is partitioned into training data and test data.
- Base models: The Level-0 models consist of RF and LR. These models utilise training data to generate aggregated predictions at level 0.
- Level-0 Predictions: Applying a base model to a set of training data generates several level-0 predictions.
- Meta Model: The stacking model incorporates only one meta-model that uses majority voting to efficiently combine the predictions of the base models. The level-1 model is also known as the meta-model.
- Level-1 Prediction: The meta-model acquires the ability to effectively amalgamate the predictions generated by the base models and is trained using the diverse predictions produced by each unique base model. Figure 1 illustrates the block diagram of the proposed stacking ensemble learning architecture.

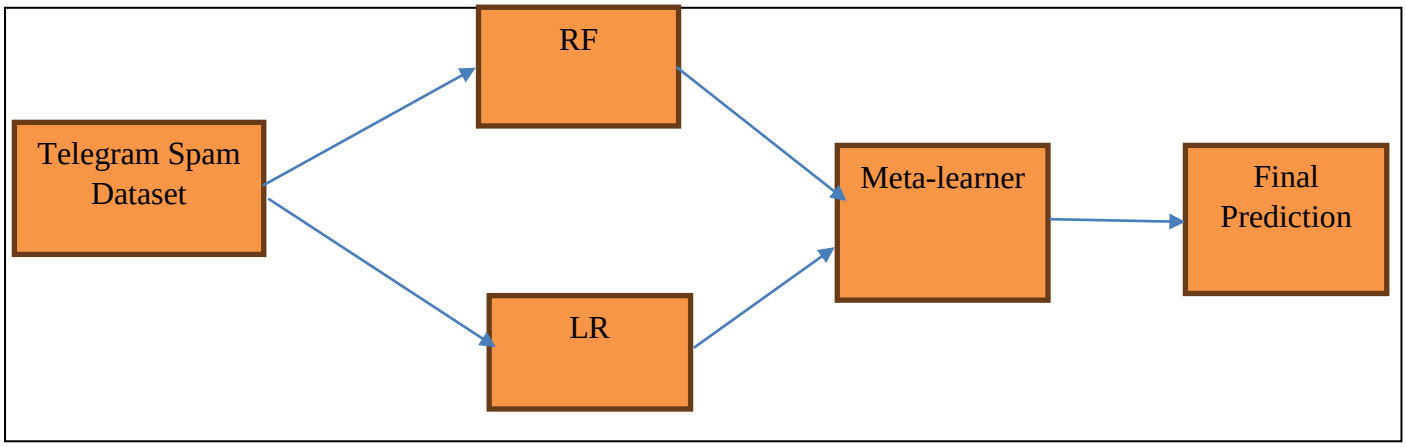

Figure 1: Block Diagram of the Proposed Ensemble Learning Architecture

A snapshot of the proposed model from the Python programming environment is shown in Figure 2.

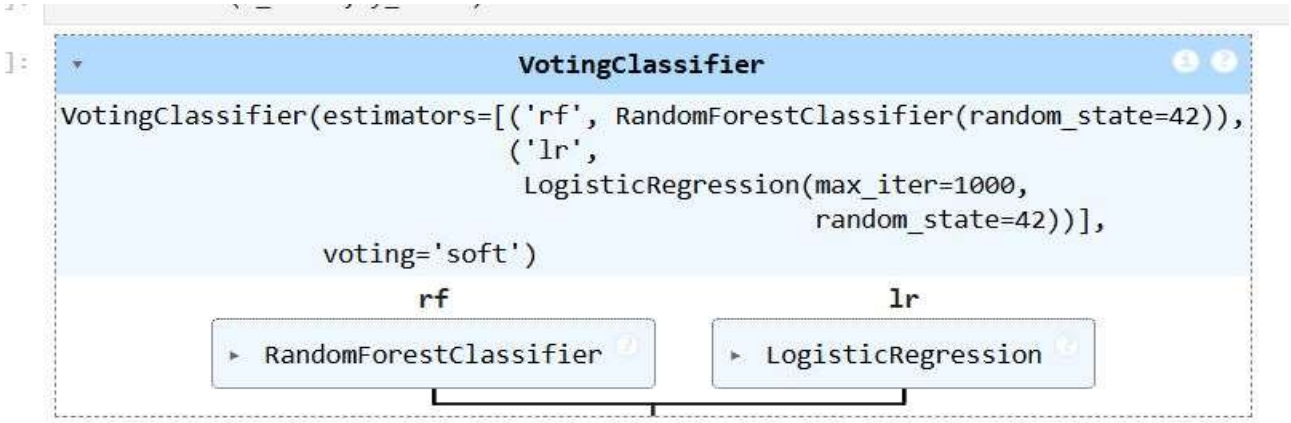

Figure 2: Snapshot of the Proposed Ensemble Model from the Python Environment

Experimental Settings

The experimental and parameter settings for the study are shown in Table 1. The process of training a model entails the selection of appropriate values for each weight and bias parameter based on labelled samples. The setting of parameters is a crucial stage in the training process of machine learning models. The parameters employed to regulate spam datasets during the training and testing phases of the models are comprehensively depicted in Table 1. These factors play a crucial role in refining the effectiveness of the model. The models were trained with the pandas, NumPy and scikit-learn tools for machine learning computation in Python programming environment.

Table 1: Experimental Settings and Parameter Tuning of RF, LR and the Ensemble Algorithm.

Model	Hyperparameter	Values
RF	n_estimators	100
	random_state	42
LR	max_iter	1000
	random_state	42
Ensemble	n_estimators	100
	random_state	42

Performance Metrics

To evaluate the performance of any ML model, we compute some values based on the comparison between the prediction results obtained from each model and the original data values. These evaluation metrics determine accuracy the model and which model is the best (Ahmed *et al.,* 2023; Bako *et al.,* 2023). We evaluated all the models under comparison by computing Accuracy, Precision, Recall and F-Measure.

Accuracy: is the ratio of summation of the right predictions (true positive + true negative) out of the total predictions (true positive, true negative, false positive, and false negative). The higher the value of Accuracy, the better the performance of a model. It is a simple straightforward method to evaluate a model's performance (Ahmed *et al.,* 2023; Bako *et al.,* 2023).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: is the ratio of the items that the model correctly classified as positive (TP) out of the whole correctly and incorrectly classified as positive (TP + FP) (Ahmed *et al.,* 2023; Bako *et al.,* 2023).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall: is the sensitivity of the model and it is the ratio of the items that the model correctly classified as positive (TP) out of the correctly classified as positive and incorrectly classified as negative (TP + FN) (Ahmed *et al.,* 2023; Bako *et al.,* 2023).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F-Measure is a value that measures the accuracy of a binary classification model performance. It is the harmonic mean H of the precision and recall (Ahmed *et al.,* 2023; Bako *et al.,* 2023).

$$\text{F-Measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

RESULTS AND DISCUSSION

This section presents the results and examines the significant discoveries derived from our simulations. The experiment was carried out in jupyter computing environment (Python 3). The Telegram spam dataset was used for the purpose of training and testing the classifiers. The effectiveness of all the models was evaluated by computing Accuracy, Precision, Recall and F-Measure.

Table 2 illustrates the performance of the models on the benchmark dataset. It is clear that the proposed ensemble model and RF algorithm perform better than the LR in all classification metrics considered on the benchmark dataset.

Table 2: Classifiers' performance on the Telegram Spam dataset

Classifier		Precision	Recall	F-Measure	Accuracy
RF		0.94	0.99	0.96	0.94
	ham	0.96	0.83	0.89	
	spam				
LR	ham	0.92	0.98	0.95	0.93
	spam	0.94	0.79	0.86	
Ensemble	ham	0.93	0.99	0.96	0.94
	spam	0.96	0.82	0.89	

The confusion matrix provides a practical way for assessing the performance of the classifiers, where each row of the table denotes actual rates of the class whereas each column indicates the predictions. The confusion matrix for the RF is shown in Figure 4. It means that the RF algorithm classifies correctly 2876 spam messages as spam, classifies wrongly 196 ham messages as spam, classifies correctly 961 ham messages as ham and classifies wrongly 37 spam messages as ham.

Figure 4: Confusion Matrix of the Random Forest Algorithm

Figure 5 illustrates the confusion matrix for the LR model. It indicates that LR algorithm identifies correctly 2854 spam messages as spam, identifies wrongly 242 ham messages as spam, identifies correctly 915 ham messages as ham and identifies wrongly spam messages as ham.

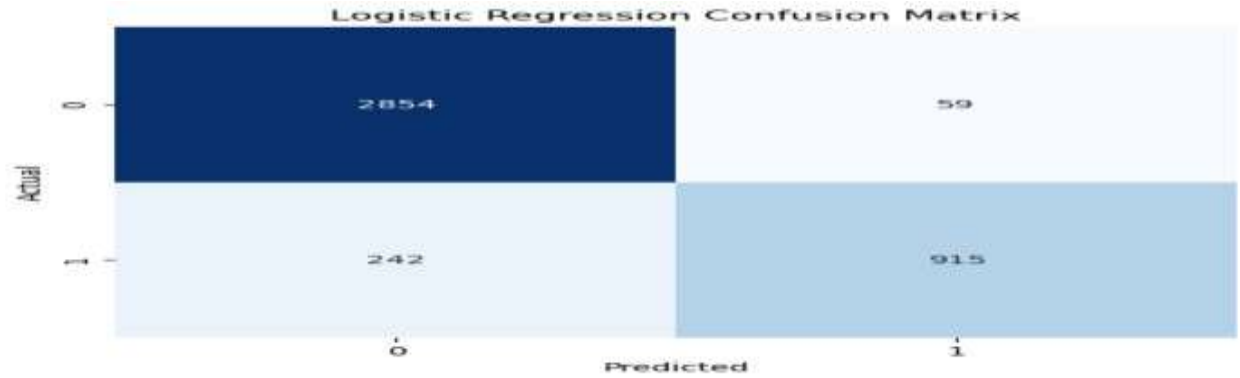

Figure 5: Confusion Matrix of the Logistic Regression

The confusion matrix for the proposed ensemble model is shown in Figure 6. The model classifies correctly 2877 spam messages as spam, classifies wrongly 210 ham messages as spam, classifies correctly 947 ham messages as ham and classifies wrongly 36 spam messages as ham.

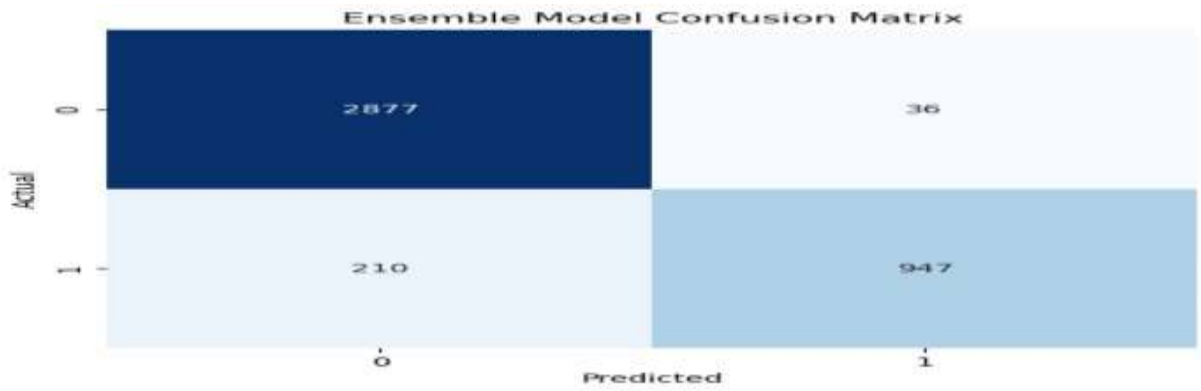

Figure 6: Confusion Matrix of the Ensemble Model

This performance analysis is also illustrated using ROC curves to provide insight into the trade-offs between sensitivity (recall) and specificity. It plots True Positive Rate (Recall) against the False Positive Rate (1- Specificity). The Area Under Curve (AUC) indicates the extent of separability and measures how good a model is at classifying between positive and negative classes. The ROC curves of the RF, LR and proposed ensemble model are depicted in Figure 7, Figure 8 and Figure 9 respectively. The models show a higher AUC which indicates better performance at predicting spam and ham messages.

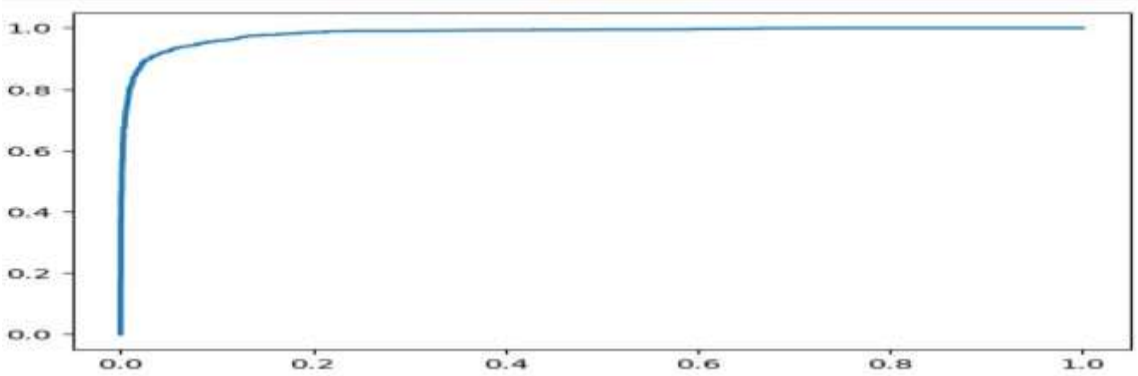

Figure 7: ROC Curve for the Random Forest

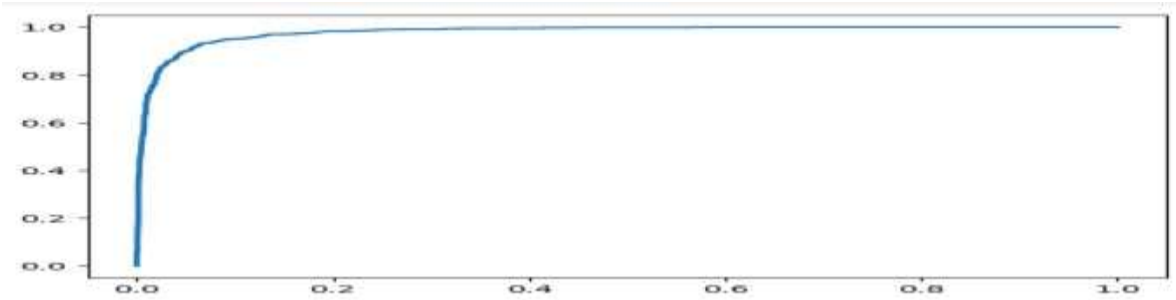

Figure 8: ROC Curve for the Logistic Regression

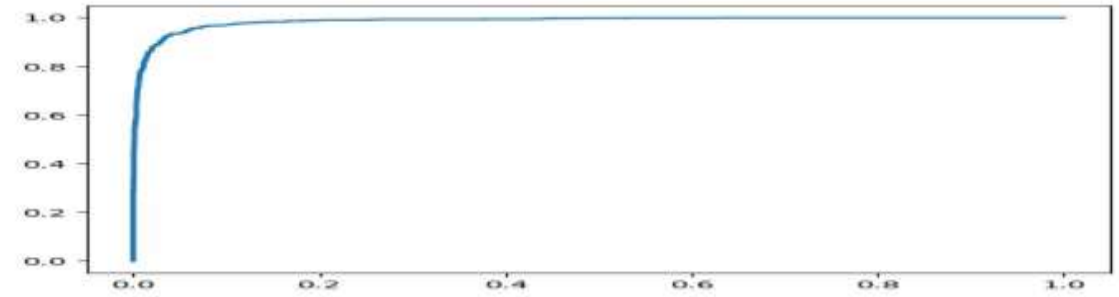

Figure 9: ROC Curve for the Ensemble Model

CONCLUSION

This research presents ensemble learning algorithm for classifying spam in Telegram messages. Telegram is a popular instant messaging platform that started in 2013. We used the Telegram spam dataset obtained from Kaggle to train and test the proposed ensemble method and the base learners (RF and LR). Experiment results show that the proposed ensemble model and RF classifier performed better than the LR model. algorithm in classifying Telegram spams in the benchmark dataset used in this study. The findings indicate that the ensemble model and RF algorithms promise to be a good approach for Telegram spam detection. Further research work should leverage more sources of data and the use of other state- of the -art machine learning techniques should be considered for larger datasets.

REFERENCES

Ahmed, R. N., Javed, A., & Bedewi, W. (2023). Is COVID-19 Being Used to Spread Malware. *SN Computer Science*, *4*(4). https://doi.org/10.1007/s42979-023-01838-6

Baaqeel, H., & Zagrouba, R. (2020, November 28). Hybrid SMS spam filtering system using machine learning techniques. *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*. https://doi.org/10.1109/ACIT50332.2020.9300071

Bako, H. S., Ambursa, F. U., Galadanci, B. S., & Garba, M. (2023). PREDICTING TIMELY GRADUATION OF POSTGRADUATE STUDENTS USING RANDOM FORESTS ENSEMBLE METHOD. *FUDMA JOURNAL OF SCIENCES*, *7*(3), 177–185. https://doi.org/10.33003/fjs-2023-0703-1773

Castano, F., Fernandez, E. F., Alaiz-Rodriguez, R., & Alegre, E. (2023). PhiKitA: Phishing Kit Attacks dataset for Phishing Websites Identification. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3268027

Dada, E. G., & Bassi, J. S. (2018). Logistic Model Tree Induction Machine Learning Technique for Email Spam Filtering. In *The Pacific Journal of Science and Technology-96* (Vol. 19, Issue 2). http://www.akamaiuniversity.us/PJST.htm

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam

filtering: review, approaches and open research problems. *Heliyon*, *5*(6). https://doi.org/10.1016/j.heliyon.2019.e01802

Dada, E. G., Birma, A. I., & Gora, A. A. (2024). Ensemble Machine Learning Algorithm for Diabetes Prediction in Maiduguri, Borno State. *Mikailalsys Journal of Mathematics and Statistics*, *2*(2), 46–73. https://doi.org/10.58578/mjms.v2i2.2875

Dada, E. G., & Joseph, S. B. (2018). Random Forests Machine Learning Technique for Email Spam Filtering. In *University of Maiduguri Faculty of Engineering Seminar Series* (Vol. 9, Issue 1).

Dada, E. G., Oyewola, D. O., & Yakubu, J. H. (2022). Arid Zone Journal of Basic and Applied Research Power Consumption Prediction in Urban Areas using Machine Learning as a Strategy towards Smart Cities. In *AJBAR* (Vol. 1, Issue 1).

Dargahi Nobari, A., Sarraf, M. H. K. M., Neshati, M., & Erfanian Daneshvar, F. (2021). Characteristics of viral messages on Telegram; The world's largest hybrid public and private messenger. *Expert Systems with Applications*, *168*. https://doi.org/10.1016/j.eswa.2020.114303

Domor Mienye, I., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.DOI

Gbenga Dada, E., Opeoluwa Oyewola, D., Bassi Joseph, S., & Baba Dauda, A. (2021). Ensemble Machine Learning Model for Software Defect Prediction. In *Adv Mach Lear Art Inte* (Vol. 2, Issue 1). www.opastonline.com

Kumar, C., Bharti, T. S., & Prakash, S. (2022). A hybrid Data-Driven framework for Spam detection in Online Social Network. *Procedia Computer Science*, *218*, 124–132. https://doi.org/10.1016/j.procs.2022.12.408

Mambina, I. S., Ndibwile, J. D., Uwimpuhwe, D., & Michael, K. F. (2024). Uncovering SMS Spam in Swahili Text Using Deep Learning Approaches. *IEEE Access*, *12*, 25164–25175. https://doi.org/10.1109/ACCESS.2024.3365193

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. In *IEEE Access* (Vol. 10, pp. 99129–99149). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3207287

Morgia, M. La, Mei, A., Mongardini, A. M., & Wu, J. (2021). *It's a Trap! Detection and Analysis of Fake Channels on Telegram*.

Oh, H. (2021). A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model. *IEEE Access*, *9*, 144121–144128. https://doi.org/10.1109/ACCESS.2021.3121508

Oyewola, D. O., & Dada, E. G. (2022). Machine Learning Methods for Predicting the Popularity of Movies. *Journal of Artificial Intelligence and Systems*, *4*(1), 65–82. https://doi.org/10.33969/ais.2022040105

Pinandito, A., Perdana, R. S., Saputra, M. C., & Az-Zahra, H. M. (2017). Spam detection framework for Android Twitter application using Naive Bayes and K-Nearest Neighbor classifiers. *ACM International Conference Proceeding Series*, 77–82. https://doi.org/10.1145/3056662.3056704

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00592-x