Classifying Swahili Smishing Attacks for Mobile Money Users: A Machine-Learning Approach

P.Siva Teja, Department of CSE, Narayana Engineering College, Gudur.

Abstract: *Due to the massive adoption of mobile money in Sub-Saharan countries, the global transaction value of mobile money exceeded $2 billion in 2021. Projections show transaction values will exceed $3 billion by the end of 2022, and Sub-Saharan Africa contributes half of the daily transactions. SMS (Short Message Service) phishing cost corporations and individuals millions of dollars annually. Spammers use Smishing (SMS Phishing) messages to trick a mobile money user into sending electronic cash to an unintended mobile wallet. Though Smishing is an incarnation of phishing, they differ in the information available and attack strategy. As a result, detecting Smishing becomes difficult. Numerous models and techniques to detect Smishing attacks have been introduced for high-resource languages, yet few target low-resource languages such as Swahili. This study proposes a machine-learning based model to classify Swahili Smishing text messages targeting mobile money users. Experimental results show a hybrid model of Extratree classifier feature selection and Random Forest using TFIDF (Term Frequency Inverse Document Frequency) vectorization yields the best model with an accuracy score of 99.86%. Results are measured against a baseline Multinomial Naïve-Bayes model. In addition, comparison with a set of other classic classifiers is also done. The model returns the lowest false positive and false negative of 2 and 4, respectively, with a Log-Loss of 0.04. A Swahili dataset with 32259 messages is used for performance evaluation.*

INDEX TERMS*:Natural language processing, mobile money, machine-learning, SMS, Sub-Saharan Africa, social engineering, smishing.*

I.INTRODUCTION

Swahili is a Bantu language native to the Swahili people. Swahili is the most widespread language south of the Sahara. Swahili is one of the official languages of the African Union (AU), Southern African Development Community (SADC), and East African Community (EAC). It is spoken by more than 16 African countries and is the lingua franca of the Indian coastal region spanning from Somalia to Mozam bique and some parts of Zambia, Malawi, South Africa, The Comoros, Botswana, and The Democratic Republic of Congo. Swahili currently borrows 30–40% of its vocabulary from non-Bantu languages, where most of the borrowings are from Arabic and Persian . Swahili continues to be the most widely spoken Bantu dialect . It is among the 10 most spoken languages in the world, with more than 200 million The associate editor coordinating the review of this manuscript and approving it for publication was Joey Tianyi Zhou. native or second-language speakers. Despite their popu larity, many of the 7000+, languages and language varieties in use today around the world do not have adequate data to warrant their processing on digital platforms . Researchers have focused more on 20 languages out of the 7000+, leaving the vast majority of languages in limbo. Hence, the terms ‘‘low-resourced’’ and ‘‘high-resourced’’ languages . Low resource can mean less studied, scarce data sources, fewer computational tools, fewer digital contents, taught locally, or low density. However, many of these low resource languages, such as Swahili, Bengali, and Punjab, are spoken by millions of people . Prior to 2006, governments in lower-middle income coun tries were perspiring over the problem of financial inclusion. Apart from urban populations, which form a fraction of the population, a large part of the population in most Sub Saharan countries has no access to formal financial services..

This work is licensed under a Creative Commons Attribution 4.0 International License

II.FUNCTIONAL OVERVIEW

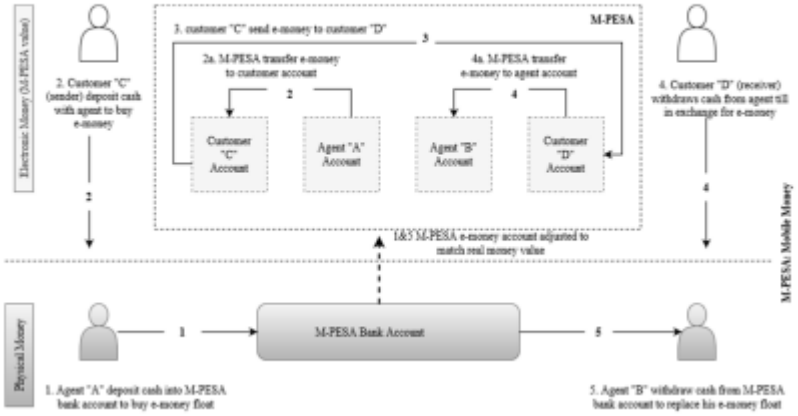

FIGURE 1. Overview of mobile money functionality a case of MPesa

On start after giving A telecommu nication company in Kenya proposed a solution that uses a mobile number as a wallet to provide financial services termed ‘‘MPesa’’ . Economides & Jeziorski in define mobile money (MPesa) services as a wallet that is associated with a mobile number and functions as a traditional bank account. Hughes and Lonie argue that the mobile money ecosystem involves mainly three actors: a customer, an agent, and a mobile network operator. Customers and agents can perform the following actions: deposit, withdrawal, sending, and receiving of cash. Mobile network operators ensure con nectivity between the other two actors. SMS allows users in the ecosystem to communicate. An overview of mobile money platform operations is presented in Fig. 1 Inspired by advancements in machine-learning techniques coupled with promising results obtained in message classifi cation. This study proposes a machine-learning based model to classify Swahili Smishing text messages targeting mobile money users. Machine-learning techniques are advantageous to other techniques as they can detect both known malware and obfuscated malware . The contributions of this study, organized and carried out under a real-world Swahili Smish ing dataset collected from mobile money users in Tanzania, are summarized as follows:

• Introducing a hybrid machine-learning model to effec tively classify Swahili Smishing messages based on the unique features these messages share

. • Evaluate the performance of the proposed model by comparing it with other traditional models classifying Smishing messages in other languages.

• We reviewed and categorized the typical existing approaches for Smishing message detection.

• We have highlighted message signatures used by social engineers during Smishing attacks aimed at mobile money users.

• The study offers a new real, non-encoded Swahili Smishing dataset for further studies.

The proposed model would save mobile money users from financial losses they incur as a result of social engineering attacks that keep on utilizing local dialects that are less studied. The rest of this paper is structured as follows: The second section will discuss related works and the objectives of this paper. The third section will elaborate on methods used to conduct the research. The fourth section will deliberate on the results of the study. The fifth section will discuss the results of the study. Lastly, the sixth section will conclude the study and give future recommendations.

This work is licensed under a Creative Commons Attribution 4.0 International License

III. RELATED WORKS

Recently, spam filtering has caught the interest of vari ous researchers around the globe due to the unprecedented increase in spam message flow on networks. The proposed work spans from detecting email spam, web spam, and spam on social networks. A variety of studies have been con ducted to investigate email spam and web spam in a wide spectrum . Researchers have also discussed various Smishing detection approaches . Over the years, Smishing detection has been dependent on blacklisting, heuristics, and visual analytics methods. For instance, Chen et al. proposed a Smishing control system based on trust management; the system aimed to control or filter Smishing based on trust relations between the sender and receiver of messages. A rule-based approach is proposed in to detect Smishing messages in a mobile environment. They identified nine rules, the majority of which had char acteristics such as bogus links, mobile numbers, advertise ments, messages with self-answering questions, the intention of fake news spreading, and lottery winning. A rule-based classification algorithm was applied and yielded a 92% true positive rate and a 99% true negative rate during evaluation. Kipkebut et al used a Naïve-Bayes algorithm to clas sify spam messages targeting mobile money users in Kenya.

The study collected spam messages written in English and used the Weka toolkit to perform the experiments. After experimentation, they managed to attain an accuracy of 96.1039%. People in Kenya use more than one language to communicate, English being one of them. However, the study did not consider messages that were written in other languages, such as Swahili, which is spoken widely in Kenya. Baek et al. propose a detection mechanism for analyzing real-time behavior via recording changes to system files. The system intends to detect unknown malware that targets IoT devices using a two-stage mechanism. However, loss of fea tures during feature vectorization and selection in stage 1, and high data and hardware needs during training of deep learning models, limits the detection performance of the proposed 2-Mad scheme. Maseno et al. proposed a vishing detection model that breaks down the process of an attack into manageable components and guidelines to aid user decision-making. This rule-based model had five rules that worked on the basis of emotion, script completeness, information requested level, and phone number. The rules were applied to the technical complexity, psychological factors, and information sensitiv ity of the attacks . Bryan presented a framework for detecting Smishing and vishing attacks related to mobile money transactions. The framework proposes what customers should do when faced with such an attack .

However, Hazarika et al. Joo & Yoon Kang et al. Lee et al. argue that Smishing techniques keep on devel oping where humans might be left in the dark with new techniques that more often than not follow the same pattern. Therefore, new countermeasures become a necessity. On the other hand, advancements in text classification techniques offer suitable and promising solutions that scale well to the current cyber environment.

This work is licensed under a Creative Commons Attribution 4.0 International License

IV. METHODS

The aim of our work is to investigate an appropriate machine-learning algorithm to classify Smishing messages targeting mobile money users. This study makes use of machine-learning models since they are less data and hard ware hungry as compared to deep learning models . Naturally, Smishing messages targeting mobile money users use words in a well-orchestrated pattern and a mobile number to receive electronic money from a victim. Fig. 2 presents the overall architecture of the proposed approach. After data collection, messages are preprocessed by removing unneces sary words such as Stopwords. Tokenization is then applied, where a list of sentences is converted into a list of words. This process is necessary since the vectorization of text happens at the words level and character level. The study considers word vectorization to minimize the dimension of the resultant vector, where words are vectorized with the help of count and TFIDF vectorizer. Feature selection and parameter turning were applied during model training. This study trained the model with two techniques; bag of words and n-gram. We use 2-5 n-grams to find the best performing model.

AUTHOR	YEAR	CLASSIFIER	DATASET	LANGUAGE
Mishra & Soni. [52]	2021	Backpropagation	Almeida *et al.* [53]	English
Liu *et al.* [35]	2021	Logistic Regression	360-Mobile safe	English
Mishra & Soni. [51]	2020	Naïve-Bayes	Almeida *et al.* [53]	English
Baaqeel & Zagrouba. [46]	2020	K-Means and SVM	UCI Machine-Learning repository	English
Saeed. [38]	2020	Discrete Hidden Markov Model	UCI Machine-Learning repository	English

TABLE 1. Review of recent Smishing detection models.

A.DATA COLLECTION

Data collection activity was conducted in Tanzania and a series of experiments were performed. Mobile network oper ators were purposely selected based on their mobile money market share. A purposive sampling technique is selected due to its ability to match the aims and objectives of the research . Out of the available users of the mobile money platform, university students were used as the selected cluster to col lect legitimate messages. According to Palinkas et al. a rather small and purposively selected sample may be included in a study with the aim of amplifying the depth as opposed to breadth of understanding. Therefore, this study collected its dataset from mobile network operators and uni versity students. The dataset is available on Github with special permission .

B. DATASET NORMALIZATION

In most cases, an imbalanced dataset signifies that there are fewer examples of a minority class in the dataset for a machine-learning algorithm to learn the decision boundary. In this particular case, the dataset is highly imbalanced as the number of unique legitimate messages is in the thousands while we managed to collect three hundred and two unique Smishing messages. One approach to balancing the dataset would be to duplicate the

This work is licensed under a Creative Commons Attribution 4.0 International License

minority class. This technique can balance the dataset but does not add any additional infor mation to the dataset for the model to learn. A different approach is to use a synthetic minority over-sampling tech nique (SMOTE). SMOTE tries to oversample the minority class by creating synthetic examples rather than oversampling with replacement. As Chawla et argue, SMOTE creates synthetic examples in a less application-specific manner by operating in feature space rather than data space. SMOTE draws a line between sample examples in the dataset that are close in feature space and, thereafter, tries to generate new examples that will be close to the feature space created.

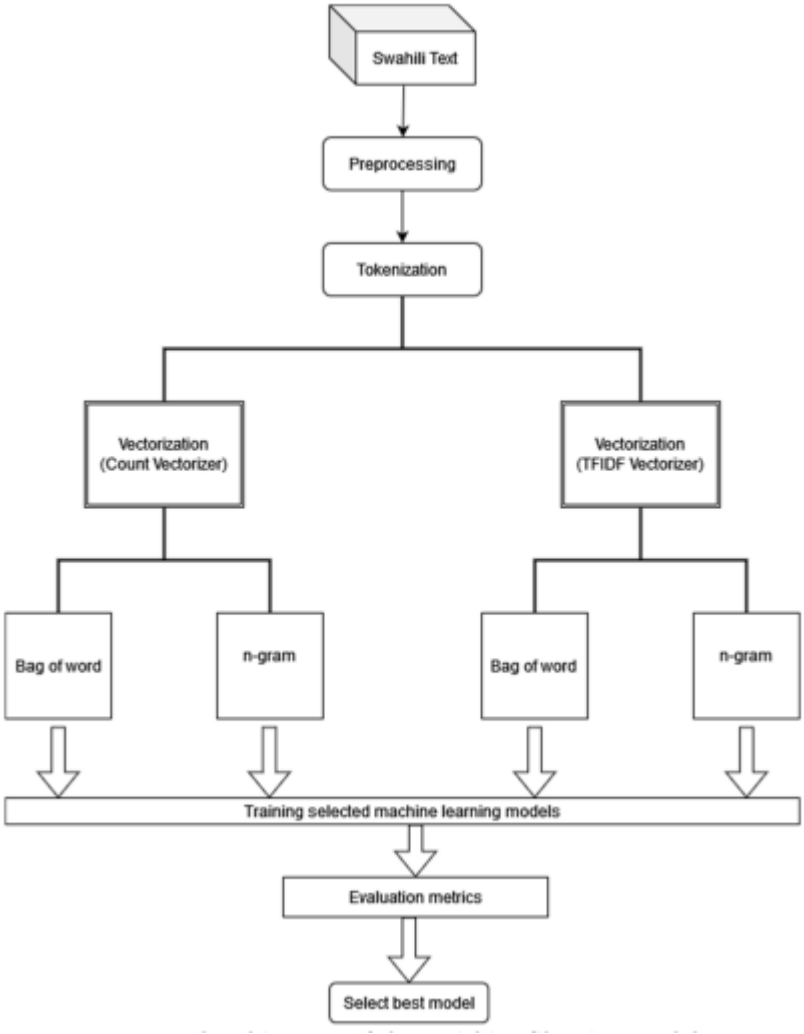

FIGURE 2. General architecture of the Smishing filtration model.

C. TEXT PREPROCESSING AND ENCODING The dataset was manually and consistently encoded by experts with spam and legitimate labels. Text preprocessing and data cleaning were done with the help of Python library functions. We converted all the contents of the dataset to lowercase characters, and punctuation marks were removed. Because of the study context, numeric values were not deleted. They can mean a figure as a lump sum to be trans ferred to another number, a way to prevent the rule-based system from identifying the messages, or

This work is licensed under a Creative Commons Attribution 4.0 International License

a mobile number that an attacker uses to receive cash. A list of Stopwords from the study by Masua & Masasi was used to remove Stopwords from the dataset. The dataset was tokenized to produce a list of words considered as input-features.

D. FEATURE SELECTION AND VECTORIZATION

The target column was encoded into ‘‘0’’ and ‘‘1’’, where all legitimate messages were encoded with label ‘‘0’’ and all Smishing messages were encoded with label ‘‘1’’. When con verting the text content of the dataset into vectors, we exper imented with two kinds of vectorization techniques. Count and TFIDF vectorization techniques were considered for this setup. The count vectorization technique uses the frequency of words in the document and creates a sparse matrix to repre sent the occurrence of each word in the document. The TFIDF vectorizer creates a vector by giving weight to frequent words in a document but rare words in the whole dataset. This creates a feature space that is better than the count vectorizer feature space. Created vectors contain individual weights of each token for further processing.

E. EVALUATION METRICS

The suggested algorithms employ a set of metrics to measure their performance. The metrics gauge the performance in terms of the percentage of correct examples detected and the number of misclassifications the algorithm makes. The study made the following assumptions: N = {A set of all documents in our corpus} NL = {A set of all document with legitimate content} NS = {A set of all documents with Smishing content} The following evaluation metrics were used to check the performance of algorithms: True Positive (TP): NS classified as NS by the algorithm. True Negative (TN): NL classified as NL by the algorithm. False Negative (FN): NL classified as NS by the algorithm.

	Predicted as Legitimate SMS	Predicted as Smishing SMS
Labeled as Legitimate SMS	True Positive (TP)	False Negative (FN)
Labeled as Smishing SMS	False Positive (FP)	True Negative (TN)

TABLE 2. Confusion matrix.

V.RESULTS

A. COMPARISON

This work is licensed under a Creative Commons Attribution 4.0 International License

It illustrates a comparison of our proposed model with various Smishing detection models. We looked at three mobile money-specific Smishing detection models and one generic Smishing detection model. The criteria for compar ison are based on the model's security measures and imple□mentation methodologies. The comparison chart clearly shows that we employed an innovative approach in our algorithm to recognize Swahili Smishing messages that target mobile money customers. Extratree feature selection and scoring of Swahili Smishing text aid in the creation of a Smishing message signature and increase the likelihood of detection. The use of rules and heuristic classification methods is used in other models, but the creation of message patterns has been difficult to generate. Since the messages that target mobile money users are very different from other Smishing messages. The proposed model has a high accuracy score compared to general Smishing detection models. High accuracy is the result of a proper Swahili dataset that we were able to collect from various stakeholders. Furthermore, a comparison with baseline models for text classification shows that baseline models do not perform well with Swahili text. A lower accuracy for the Swahili dataset can be attributed to the fact that the formation of words and sentences in the Swahili language is very different from other well studied languages such as English, which has been extensively used by other researchers.

B. MESSAGE LENGTH

Messages were inspected and it was found that legitimate messages are usually short, with a mean value of forty-nine (49) words per message, while Smishing messages have a mean value of one hundred and twenty-six (126) words per message, as depicted in Fig. 3 and Fig. 4

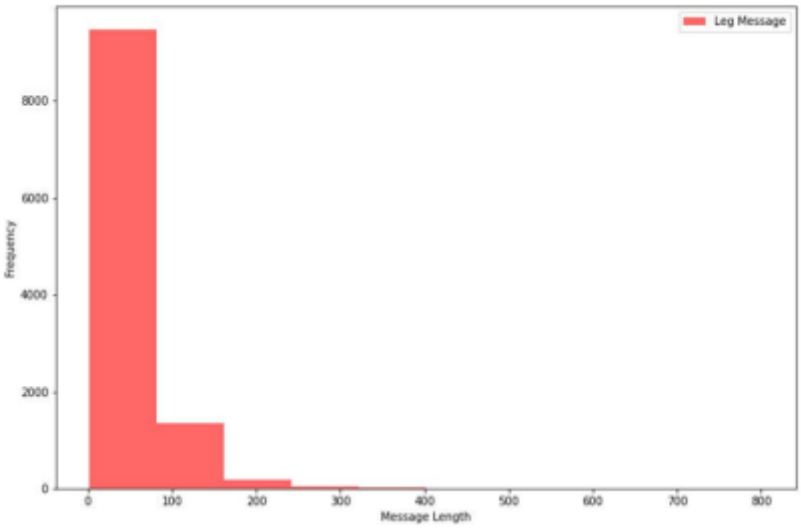

FIGURE 3. Length of legitimate messages.

This work is licensed under a Creative Commons Attribution 4.0 International License

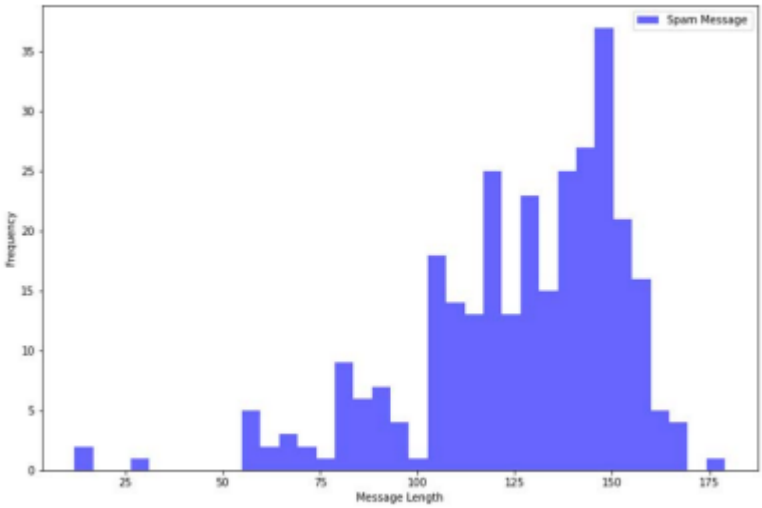

FIGURE 4. Length of Smishing messages.

C. FREQUENT WORDS

The study examined the dataset for the most frequently used words in legitimate messages as opposed to Smish ing messages. A word-cloud is printed to show frequently used words, which are further considered features of our model to differentiate Smishing from legitimate messages. Fig. 5 shows the most prominent words used in Smish ing messages, such as pesa (translates to money), namba (phone number), tiba asili (traditional medicine), and piga sim (phone call). Smishing messages seem to contain content that requires a user to call or send cash to an unknown mobile number. On the other hand, displays the top words used in legitimate messages. Legitimate messages contain words such as watu (translates to people), mzee (an old person), leo (today), nchi (a country), ndio/sawa (agreeing), mama (mother). These are normal words that have nothing to do with money or transactions. Some of the words that are more frequent in messages appear darker and with a larger font on the word-cloud than less frequent words. For example, ‘‘ndo’’ is the most frequently used legitimate word, while ‘‘namba’’ is at the top of the spam word list.

D. TOP FEATURES

Among all 21,408 features, we show the top twenty features of our dataset in Fig. 7. As it can be seen from Figure 5, the word piga (which translates to ‘‘call a number’’) has the highest importance since, most of the time, attackers would require a user to call a number that is present in the Smishing message. It’s closely followed by the word litakuja (preordination). This word is used mostly because it’s an authentication check procedure while transferring cash. Such FIGURE 5. Word-cloud of legitimate messages. FIGURE 5. Word-cloud of Smishing messages. FIGURE 7. Top 20 features from Smishing

This work is licensed under a Creative Commons Attribution 4.0 International License

dataset. words are used as message signatures by the model to increase the likelihood of Smishing messages detection.

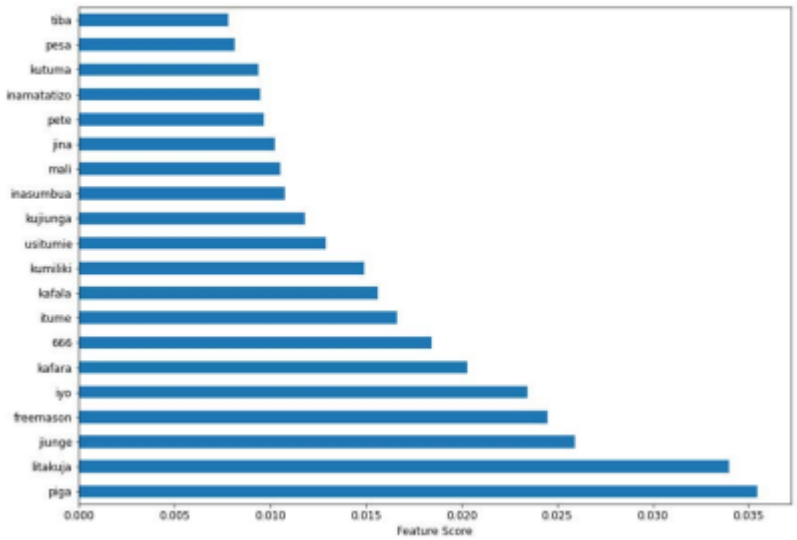

FIGURE 5. Top 20 features from Smishing dataset.

E. MODEL PERFORMANCE

Table 3 shows the performance of various models that are known to have performed well on binary classification tasks:

MODEL	TRAINING TIME	ACCURACY	AUC	F1-SCORE	LOG-LOSS
Multinomial Naïve-Bayes	6.59 ms	0.9025	0.9024	0.9099	3.38
Logistic Regression	3.77 s	0.9482	0.9528	0.9513	1.61
SVM	2.95 s	0.9473	0.9530	0.9510	1.62
KNN	1 s	0.9421	0.9519	0.9501	1.65
Random Forest	150 ms	0.9486	0.9524	0.9507	1.63
Adaboost	3.01 s	0.9468	0.9467	0.9451	1.83
Extra Tree Classifier	2.03 s	0.9547	0.9530	0.9514	1.61

TABLE 3. Model performance with count vectorizer taking 750 feature set

This work is licensed under a Creative Commons Attribution 4.0 International License

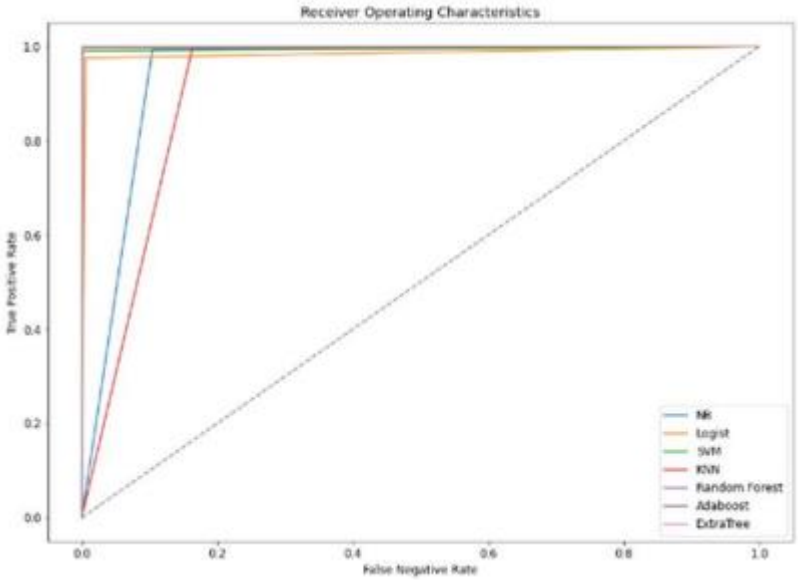

FIGURE 6. Receiver operating characteristics comparing all models.

VI.CONCLUSION

Recently, mobile network operators have seen a steep rise in Smishing attacks. These attacks can be general or targeted, with governments in the East African region pushing for financial inclusion through mobile money. Smishing attacks targeting mobile money users are skyrocketing. Hence, this paper focused on investigating an appropriate algorithm to classify legitimate messages from Smishing messages tar geting mobile money users. We successfully investigated various machine-learning algorithms to find what best fits the context in question. The results from the experiments show that Random Forest evaluates the best accuracy score of 99.86%. Therefore, it can be concluded that a hybrid of the Extratree classifier feature selection technique in conjunction with Random Forest, taking 750 as the maximum number of features vectorized by the TFIDF technique, returns the best accuracy score. In the future, we shall design a mobile application that uses the identified algorithm.

Furthermore, a deep learning methodological approach will be explored. The approach may further reduce the number of false positives and false negatives, which could be very costly to users. They could either incur financial loss or ignore an important message.

This work is licensed under a Creative Commons Attribution 4.0 International License

VI. REFERENCES

[1] A. Y. Lodhi, Oriental Influences in Swahili. A Study in Language and Cul tural Contacts. Gothenburg, Sweden: Acta Universitatis Gothoburgensis, 2000.

[2] B. E. Coleman, ‘‘A history of Swahili,’’ Black Scholar, vol. 2, no. 6, pp. 13–25, 1971.

[3] UNESCO. (2021). World Kiswahili Language Day. 41st Session, Paris. Accessed: Jan. 29, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000379702

[4] S. M. Lakew, M. Negri, and M. Turchi, ‘‘Low resource neural machine translation: A benchmark for five African languages,’’ 2020, arXiv:2003.14402.

[5] A. Magueresse, V. Carles, and E. Heetderks, ‘‘Low-resource languages: A review of past work and future challenges,’’ 2020, arXiv:2006.07264.

[6] J. Hirschberg and C. D. Manning, ‘‘Advances in natural language process ing,’’ Science, vol. 349, no. 6245, pp. 261–266, Jul. 2015.

[7] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey, ‘‘Selection criteria for low resource language programs,’’ in Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC), 2016, pp. 4543–4549.

[8] A. K. Singh, ‘‘Natural language processing for less privileged languages: Where do we come from? Where are we going?’’ in Proc. IJCNLP, 2008, pp. 7–12.

[9] Y. Tsvetkov, ‘‘Opportunities and challenges in working with low-resource languages,’’ Ph.D. dissertation, Language Technol. Inst., Pittsburgh, PA, USA, 2017.

[10] A. A. Amidu, ‘‘Kiswahili: People, language, literature and lingua franca,’’ Nordic J. Afr. Stud., vol. 4, no. 1, pp. 104–123, 1995.

[11] G. De Pauw and G.-M. De Schryver, ‘‘Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes,’’ Lexikos, vol. 18, pp. 11–13, Oct. 2011.

[12] N. Hughes and S. Lonie, ‘‘M-PESA: Mobile money for the ‘unbanked’ turning cellphones into 24-hour tellers in Kenya,’’ Innov., Technol., Gover nance, Globalization, vol. 2, nos. 1–2, pp. 63–81, Apr. 20

This work is licensed under a Creative Commons Attribution 4.0 International License