ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 6.551

+91 9940572462 +91 9940572462 ijarasem@gmail.com www.ijarasem.com

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

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

Mrs. Pranamita Nanda [1], Rahul Dev J A[2], Udhay Kumar M[3], Somanath Das R [4], Ashwin N P[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[2]Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[3]Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[4]Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

[5]Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

ABSTRACT: Due to the massive adoption of mobile money in Sub-Saharan countries, the global transaction value of mobile money exceeds. Spammers use Smishing (SMS Phishing) messages to trick a mobile money user into sending electroniccash to an unintended mobile wallet. As a result, detecting Smishing becomesdifficult. 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 classifier feature selection and Random Forest using TFIDF (Term Frequency Inverse Document Frequency) vectorization yieldsthe Multinomial Naïve-Bayes model. The model returns the lowest false positive and false negative of 2 and 4, respectively, with a Log-Loss of 0.04. ASwahili dataset with 32259 messages is used for performance evaluation. numerous models and techniques to detect Smishing attacks have been introduced for high-resource languages, yet few target low-resource languagessuch as Swahili. This study proposes a machine-learning based model to classifySwahili Smishing text messages targeting mobile money users.

KEYWORDS: Global transaction value, TFIDF, Phishing.

I. INTRODUCTION

Predict the Swahili smishing attack. Mobile money platform evolution could be attributed to the bureaucracy of owning a bank account, phishing attack on mobile user. This study proposes a machine-learning based model toclassify Swahili phishing text messages targeting mobile money users.

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 languagessuch as Swahili. This study proposes a machine-learning based model to classifySwahili Smishing text messages targeting mobile money users.

The Problem is caused by the overconfidence of users, a belief that those who fall for social engineering attacks are idiots, and rapidly changing attack vectors.

Over the years, mobile company operators have employed many ways to detect malicious text messages with little success. There are set of rules againstevery SMS going through an SMS gateway. Blacklist and whitelist techniques have also been employed to no available, because attackers keep on changing mobile numbers every now and then. Furthermore, blacklist and whitelist datasets are incapable of detecting zero-hour attacks and quickly become overpopulated and obsolete. User awareness programs on security good practice have not produced the desired results and are unlikely to reduce this vulnerability to zero.

II. RELATED WORKS

Inspired by advancements in machine-learning techniques coupled with promising results obtained in message classification. This study proposes a machine-learning based model to classify 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 conducted under a real.

The most recent major version of Python is Python 3, which we shall be using in this tutorial. However, Python 2, although not being updated with anything other than security updates, is still quite popular. Python 2.0 was released in 2000, and the 2.x versions were the prevalent releases until December 2008. At that time, the development team made the decision to release version 3.0, which contained a few small but significant changes that were not backward compatible with the 2.x versions. Pythons 2 and 3 are similar,and some features of Python 3 have been backported to Python 2. But in general,they remain not compatible.

Both Python 2 and 3 have continued to be maintained and developed, with periodic release updates for both. As of this writing, the most recent versions available are 2.7.15 and 3.6.5. However, an official End of Life date of January 1, 2020 has been established for Python 2, after which time it will no longer be maintained. Python is still maintained by a core development team atthe Institute, and Guido is still in charge, having been given the title of BDFL (Benevolent Dictator For Life) by the Python community. The name Python, by the way, derives not from the snake, but from the British comedy troupe MontyPython's Flying Circus, of which Guido was, and presumably still is, a fan. It is common to find references to Monty Python sketches and movies scattered throughout the Python documentation.

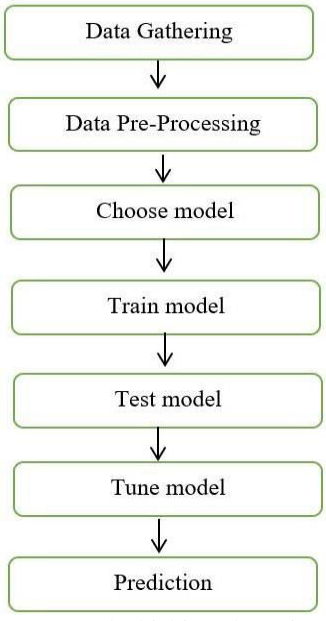
It is possible to write Python in an Integrated Development Environment, such as Thonny, PyCharm, NetBeans or Eclipse which are particularly useful when managing larger collections of Python files.

III.PROPOSED METHOD

Inspired by advancements in machine-learning techniques coupled with promising results obtained in message classification. This study proposes amachine-learning based model to classify 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 Smishing dataset collected from mobile money users. The proposed model would save mobile money users from financial losses they incur because of socialengineering attacks that keep on utilizing local dialects that are less studied.

Data Description

The Dataset collected from Kaggle.com. Two publicly available datasetsconsisting of phishing URLs and benign 'URLs



are used to evaluate the performance of the proposed phishing detection architecture. The experimental datasets are obtained from data corpus consisting of 40,000 unique phishing URLs from Phish Tank Dataset (2008) and 1,000,000,000 legitimate URLs from Alexa Experiment is conducted on the dataset of previous detection used in this study of labelled dataset, which is a revised version of the well-known dataset. The dataset is for intrusion detection

research, and it consists of tendistinct categories of six nominal feature, two binary feature and two numerical features.

Attribute	Features
Crime	Nominal Feature
Gender	Nominal Feature
Age	Numerical Feature
Income	Numerical Feature
Job	Nominal Feature
Marital Status	Nominal Feature
Education	Nominal Feature
Harm	Binary Feature
Attack	Binary Feature
Attack Method	Nominal Feature

Table 1. Data Description

Data Pre-Processing:

The dataset was manually and consistently encoded by experts with spam andlegitimate labels. Text pre-processing 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 thecontext, numeric values were not deleted. They can mean a figure as a lump sum to be transferred to another number, a way to prevent the rule-based system from identifying the messages, or a mobile number that an attacker uses to receive cash. A list of Stopwords was used to remove Stopwords from the dataset. The dataset was tokenized to produce a list of words considered asinput-features.

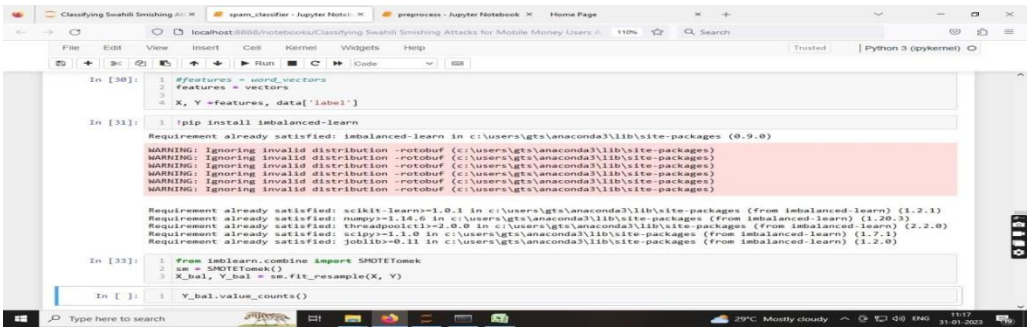


Fig 1: Data pre processing of all the dataset being evaluated.

Data Visualization:

Data visualization is the process of creating visual representations of data, such as graphs, charts, and maps, to help people better understand andanalyse the data. In simple terms, data visualization makes it easier to see andunderstand large amounts of data by presenting it in a visual format. By
creating visual representations of data, people can more easily understand and analyze complex information, leading to better decision-making and moreinformed actions. Overall, data visualization is an important tool for making sense of large amounts of data and communicating insights to others in a clearand effective way. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your owndata.

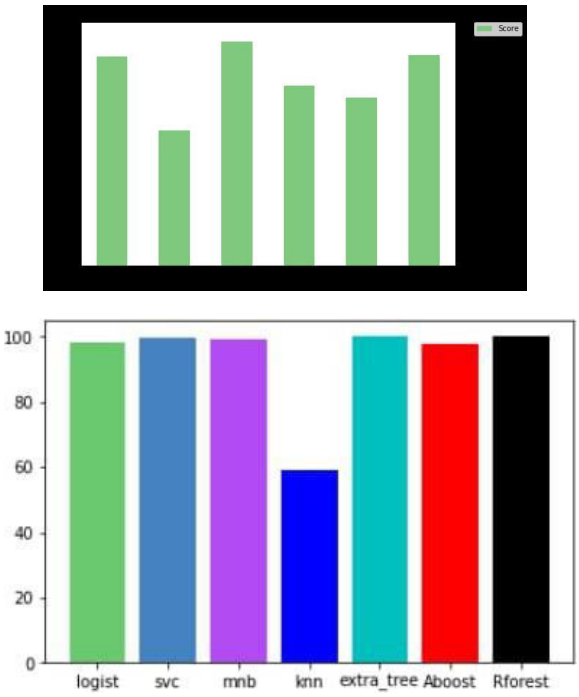


Fig 2: Types of Data Visualisation

Implementation of Algorithm:

After the Data pre-processing Techniques, the model is implemented by six chosen ML Algorithms are Multinominal Navie's Bayes; Logistics Regression; Support Vector Machine; K-nearest Neighbours; Random Forest; Adaboost; ExtraTreeClassifier. Among the six chosen models, on a feature vector created by count vectorizer, while Multinomial Naïve-Bayes performed poorly, the task of classifying Smishing messages is a sensitive one, and the return of false positives and false negatives should be taken into account.

Support Vector Machine Classifier:

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problemsas well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. SVM works by mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable.

Adaboost Algorithm:

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. It works on the principle of learners growingsequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boostingwith a slight difference.

Multinominal Naive's Bayes Algorithm:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes

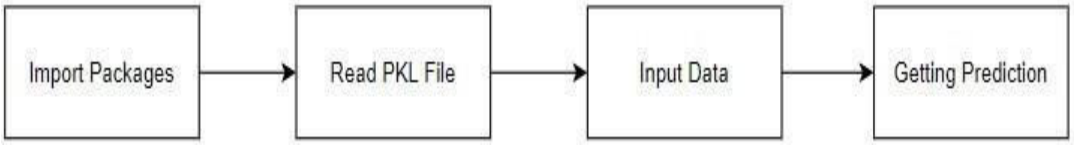
classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Deployment:

In this module the trained machine learning model is converted into pickle data format file (.pkl file) which is then deployed for providing better user interface and predicting the output of Human Stress and Deployment used here is Django Web Framework.

Django is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries.

MODULE DIAGRAM:



IV. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set of higher accuracy score algorithm will be find out. The founded one is used in the application which can help to find the type of intrusions.

REFERENCES

[1] A. Y. Lodhi, Oriental Influences in Swahili. A Study in Language and Cultural Contacts. 2000.
[2] B. E. Coleman, "A history of Swahili," Black Sch., vol. 2, no. 6, pp. 13–25, 1971.
[3] UNESCO, "World Kiswahili Language Day," in 41st Session, Paris, 2021, vol. 41 C/61. 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," ArXiv Prepr.
ArXiv200314402, 2020.

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com