EVALUATION OF MACHINE LEARNING CLASSIFICATION MODELS FOR DETECTING ELECTRONIC FUND TRANSFERS SCAM SMSes

A MINI THESIS SUBMITTED IN PARTIAL FULFILMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

OF

THE UNIVERSITY OF NAMIBIA

BY

FILLEMON SIMANEKA ENKONO

(200936344)

September 2020

MAIN SUPERVISOR: DR NALINA SURESH (SCHOOL OF COMPUTING, UNIVERSITY OF NAMIBIA)

ABSTRACT

The last decade saw the emergence of mobile banking and a pervasive transcendence of spams from email to SMS communications. M-banking offers the users an ability to execute EFT transactions using mobile devices and allow them to receive SMS notifications acknowledging their transactions. While this provide convenience to m-banking users, in the wake of SMS spams it also presented vulnerabilities that could be exploited to scam money and goods from them. To execute these scams, spammers send forged EFT (e.g. e-wallet) deposit notification SMSes to unsuspecting users, then contact and request them to do EFT payments as refunds for the supposed erroneous deposits acknowledged by the bogus notifications. Similarly, during goods exchange, scammers use forged deposit notification SMSes to trick sellers to believe that they have paid for the goods. In Namibia, the high affordability of SIM cards and the readily available access to m-banking accounts such as e-wallet by anyone with a valid SIM number provides a favourable operating environment for the EFT SMS scammers. Inferences from literatures on novel spam filtering techniques suggested that implementing machine learning classification could help address the EFT SMS scams problem, partly motivating this study to evaluate such application. Prevalent reporting of EFT SMS scams in local media (which mostly involves the country's largest bank by market share, FNB) and the observed lack of dedicated IT solutions to address such problem were other factors that inspired this work. The study collected a dataset of ham and EFT scam SMSes, from which machine learning features for classifying SMSes were extracted. This was followed by a pre-evaluation to determine the features that allow ham and EFT scam SMSes to be classified optimally. SMSes comprising the collected dataset were then represented using the optimal features and used to train and evaluate Support Vector Machine, Naïve Bayes and Random Forest classifiers.

The evaluation results revealed that the SVM classifier was the most effective with respect to detecting EFT scam SMSes, achieving a FNR=0.00, CA=0.992, Recall=1.0 and F1-measure=0.995. The RF classifier followed with FNR=0.011, CA=0.983, Recall=0.989 and F1-measure=0.989; while the NB classifier came last with FNR=0.027, CA=0.975, Recall=0.973 and F1-measure=0.983. The envisaged future work will look to use the methods, findings and conclusions drawn in this study to guide development of mobile application(s) that implement machine learning classification to detect EFT scam SMSes.

LIST OF PUBLICATION

Enkono F. S., & Suresh N. (2020). Application of Machine Learning Classification to Detect Fraudulent E-wallet Deposit Notification SMSes. *The African Journal of Information and Communication (AJIC), 25, 1-13.*

https://doi.org/10.23962/10539/29195

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

1D-TP:	One Dimensional Ternary Patterns
BoW:	Bag of Words
CA:	Classification Accuracy
CV:	Cross Validation
EFT:	Electronic Funds Transfer
FN:	False Negatives
FNB:	First National Bank
FNR:	False Negatives Rate
FP:	False Positives
FPR:	False Positives Rate
IG:	Information Gain
IG_{tshld}:	Information Gain Threshold
KNN:	K-Nearest Neighbour
NB:	Naïve Bayes
NUS:	National University of Singapore
ROC:	Receiver Operator Characteristics
SIM:	Subscriber Identifier Module
SVM:	Support Vector Machine
RF:	Random Forest
TF:	Term Frequency
TF-IDF:	Term Frequency-Inverse Document Frequency
TN:	True Negatives
TP:	True Positives
Tp:	Ternary Pattern

UCI:	University of California Irvine
UK:	United Kingdom
URL:	Uniform Resource Locator
UTF-8:	8-bits Unicode Transformation Format
WEKA:	Waikato Environment for Knowledge Analysis

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr Nalina Suresh for the consistent encouragement, support and guidance she rendered me throughout the course of this research. May the Almighty grant her more strength and motivation to continue supporting and guiding other scholars.

Additionally, I would like to acknowledge and thank Mr Andreas Gustav (Digital Banking Channels Technical Manager at FNB Namibia), for his useful insights on the EFT SMS scams problem and his words of inspiration that encouraged me to continue with this study. I am equally grateful to all the volunteers that contributed SMSes to the corpus used in this research. Without their selfless acts, completing this work would have been impossible.

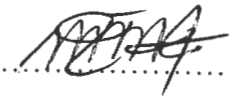
Last, but not least I would like to thank the School of Computing staff at the University of Namibia, my classmates, friends, family and everyone else that directly or indirectly assisted or contributed towards the completion of this work.

DEDICATION

I am dedicating this work to my late mother. "Our resolve to dream and fight on in the face of adversity defines who we become" indeed, continue resting in peace mom.

DECLARATIONS

I, Fillemon S. Enkono, hereby declare that this study is my own work and is a true reflection of my research and that this work or any part thereof has not been submitted for a degree at any other institution.

Fillemon S. Enkono		06/11/2020
Name of Student	Signature	Date

1. INTRODUCTION

This chapter provides the research overview. It gives a background of the research topic, explains the research problem and presents the research objectives. Furthermore, the chapter outlined the research limitations and delimitations, before concluding with an outlook for the rest of the paper.

1.1 Background of the Study

In electronic communications, text messages can be categorised into two classes, a class for desired and legitimate messages often referred as ham, and a class for unsolicited messages commonly called spams (Yan, Zhang, Kantola, & Chen, 2015). Spams comprises of phishing, advertisements and scam messages.

Ham messages often benefit both senders and receivers while spams tends to only benefit the former, frequently at a cost to the receivers (Mahmoud & Mahfouz, 2012). The one-sided benefit of sending spam messages have necessitated for detection filtering applications to allow message recipients to reduce the nuisance and losses inflicted by spams. Text message classification, which entails grouping messages into classes, constitutes a fundamental component of spams detection and filtering applications.

In the wake of the 21st century, spams and the associated detection and filtering were largely common in email communications which by then were a more affordable and pervasively used form of electronic text communication (Cormack, 2008; Khorsi, 2007). As the century progressed, evolving mobile telecommunication technologies stimulated the emergence of Short Message Service (SMS) as another predominant

form of electronic text communication. This stirred a transcendence of spams from email to SMS communications (Almeida, Gómez, & Yamakami, 2011; Hidalgo, Bringas, Sánz, & García, 2006).

Contemporaneous to the rise of mobile spams, the 21st century also saw a steady evolution and adoption of mobile banking (m-banking) services (Shaikh & Karjaluoto, 2015). Among other capabilities, m-banking offers users an ability to execute electronic fund transfers (EFT) using mobile devices and allow them to receive SMS notifications from banks, acknowledging their transactions. While such capability provides banking convenience to m-banking service users, in the wake of SMS spams, it also presented vulnerabilities that could be exploited to scam them. This culminated in the emergence of scams, where spammers utilise forged EFT deposit notification SMSes to swindle money and goods from m-banking users.

A prior literature survey could not identify an established collective name used to reference these types of scams. Therefore, borrowing from the manner in which they are committed, this study refers to them as EFT SMS scams. The literatures indicates that novel spam detection and filtering techniques are increasingly becoming centred around the application of machine learning classification (Akbari & Sajedi, 2015; Choudhary & Jain, 2017). Machine learning classification allow the spam detecting and filtering applications an ability to learn, evolve and adapt in order to remain effective as the spam behaviours changes over time (Aragão, Frigieri, Ynoguti, & Paiva, 2016).

One of the key observations from literatures is that scholarly works on machine learning SMS classification appears to mainly concentrate on detecting either the general spam SMSes or phishing SMSes (Abdulhamid et al., 2017; Akinyelu & Adewumi, 2014; Ezpeleta, Garitano, Zurutuza, & Hidalgo, 2017). This leaves out similar problems such as those caused by EFT SMS scams to continue spreading unabated as revealed in the next section. The aforementioned observation inspired this research to train machine learning models for classifying ham and EFT scam SMSes and evaluate their performances focussing on the detection of EFT scam SMSes.

1.2 Problem Statement

According to news media reports, m-banking service users across the globe suffer huge losses in finances and goods at hands of EFT SMS scammers (Arde, 2012; Christopher & Kar, 2018; Nagel, 2015). Local news media further ascertains that Namibia is not an exception to the EFT SMS scams, with most scam reports appearing to involve the country's largest bank by market share, First National Bank (Erongo, 2016). The reported EFT SMS scams also seem to predominantly involve e-wallet transactions. Based on the abovementioned reasons, this study focused on EFT SMS scams involving e-wallet transactions and collected data from FNB Namibia m-banking service users. The next paragraph explains the EFT SMS scams.

An EFT SMS scam often starts with a scammer forging a notification SMS resembling legitimate SMSes used by the bank to acknowledge EFT deposits. The scammers would then send the forged SMS to a victim and attempts to swindle either money or goods from them. Figure 1.1 describes a typical scenario of how scammers use forged

EFT scam SMSes to swindle money from m-banking users. Descriptions in the figure numbered 1 to 6 explains the scam progression.

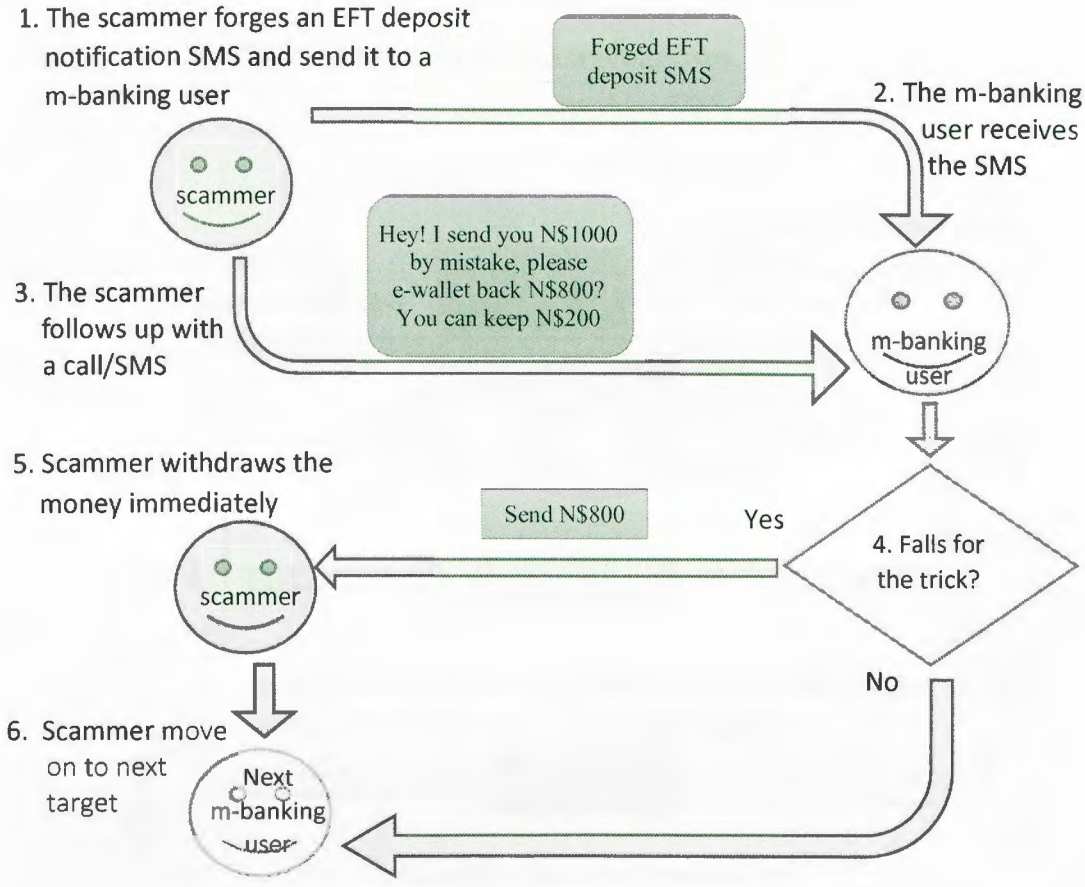


Figure 1.1: Use of EFT scam SMS to swindle money from m-banking users

When the scammers are contacting their targeted victims to request back the amount they purport to have wrongly sent, they often only request back a portion of that amount. This often entices the victims, tricking them to fall for the scams. At times, scammers resort to use social engineering to bait the victims. The screenshots of Facebook posts by EFT SMS scam victims in the latter Figure 1.3 of this section confirms the aforementioned assertions.

As cited earlier, EFT scam SMSes are not only used by spammers to swindle money from m-banking users but goods as well. A common scenario portraying how scammers use EFT scam SMSes to swindle goods from unsuspecting salespersons that are ready to accept EFT as means of payment is described in Figure 1.2. The numbered description in the figure shows how the scam progresses.

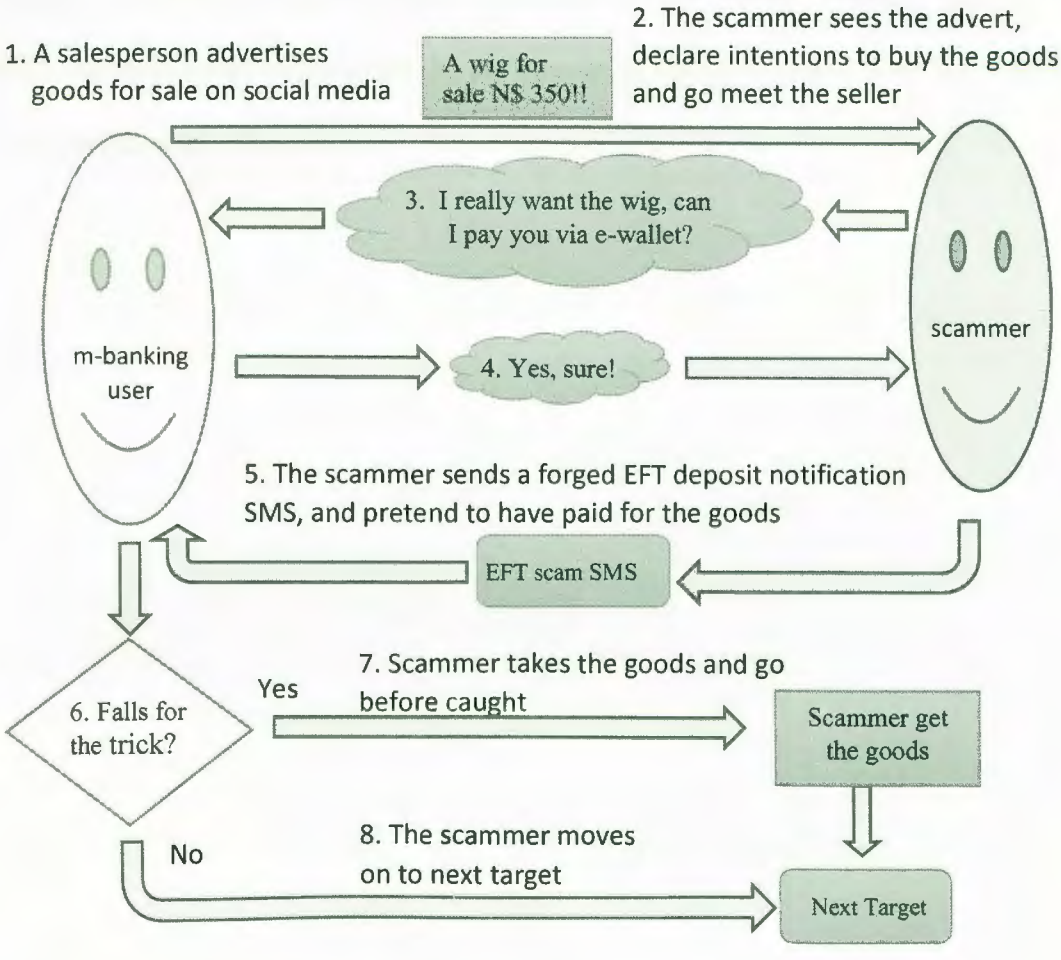


Figure 1.2: Use of EFT scam SMSes to swindle goods from salespersons

A combination of various factors provides a favourable operating environment for EFT SMS scammers, allowing them to get away with their heinous crimes with ease. One of these factors is the readily available access to m-banking accounts such as e-wallet by anyone with a valid Subscriber Identifier Module (SIM) number. Other factors are

low SIM card costs and non-requirement for users to register them before use, which allow scammers to purchase SIM numbers on a go and dispose the cards after conning people. The availability of ATMs also allows scammers to promptly withdraw the funds sent by their victims, before the victims realise that that have been conned and inform banks to reverse the transactions.

Figure 1.3 presents samples of text extracts from various media describing the EFT SMS scams in order to highlight the prevalence of such scams and disclose how they are often committed.

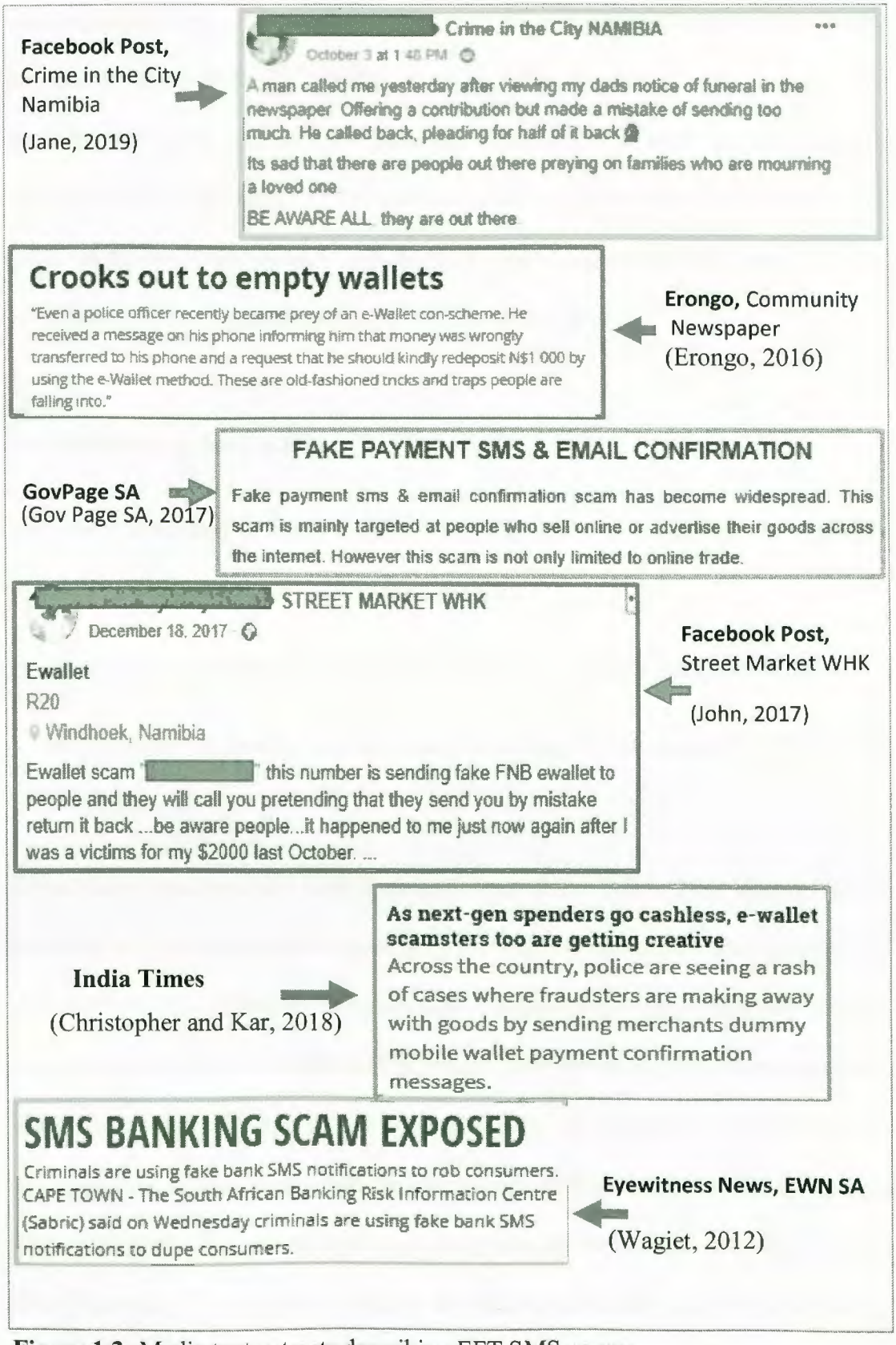


Figure 1.3: Media text extracts describing EFT SMS scams

Banking institutions often employ measures such conducting information campaigns to educate users about EFT SMS scams in order to combat such frauds. User complacencies, however create a room for these scams to continue growing. The sustained regular reporting of EFT SMS scams in local news media and the observed lack of dedicated IT solutions to address this problem, especially on the users' side has contributed to the inspirations to carry out this research.

1.3 Objectives of the Study

The study undertook to fulfil the following objectives in an attempt to contribute toward a solution to the EFT SMS scams problem:

a) Train machine learning models to classify ham and EFT scam SMSes.

b) Evaluate the models' performance prioritising on the detection of EFT scam SMSes.

Objective a) involved using SMS features to learn Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) classifiers to predict the ham and the EFT scam SMS classes. Completion of tasks such as acquisition of appropriate SMS corpus, determining a method to define and represent SMS instances, feature extraction and selection is required prior to the commencement of experiments to address this objective. These tasks along with the discussions of the considered classifiers from which NB, SVM and RF were selected are presented in the next two chapters. The final objective, b) involved evaluating the trained classifier models' performances using folds of testing data. The evaluation used several metrics (discussed in the next chapter) and paid a particular attention to the detection of EFT scam SMSes

1.4 Significance of the Study

The evaluation of the trained classifier models helped determine which classifier could be implemented into an effective practical application for detecting EFT scam SMSes on mobile phones, essentially contributing toward a solution to the EFT SMS scams problem. These contributions look set to benefit both the m-banking service users and the banking institutions. Furthermore, prospective researchers on the subject of machine learning application to counter SMS scams might draw important insights from this research.

1.5 Limitation of the Study

Collecting SMSes required for the study, particularly the EFT scam SMSes turned out to be very slow than anticipated. This limitation was addressed by extending the SMS collection time from the initial envisaged four weeks to nearly twenty weeks. This ensured that a relatively sizeable corpus of EFT scam SMSes was obtained. Furthermore, this study is based on quantitative analysis only.

1.6 Delimitation of the Study

SMSes that constituted the used ham and EFT scams corpora were collected from Namibian residents only as they were easier to reach. Moreover, as a measure to keep the scope manageable, the study only collected EFT scam SMSes involving FNB Namibia, which as highlighted in the Background of the Study section appears to be more affected by EFT SMS scams than other local banks.

1.7 Definition of Terms

Classification: Refers to the process of predicting the SMS classes.

EFT scam SMSes: Refers to forged or fake EFT deposit notification SMSes that scammers use to swindle money or goods from m-banking users.

EFT SMS scams: Refers to scams that involves the use forged EFT deposit notification SMSes.

Ham SMSes: Refers to legitimate and desired SMSes.

SMS corpus: Refers to a collection or a dataset of SMSes.

Spam SMSes: Refers to the unsolicited SMSes.

WEKA: An open source software containing a collection of machine learning algorithms for data mining tasks developed by the University of Waikato, New Zealand.

1.8 Outline of the Thesis

This thesis is structured in six chapters briefly described as follows:

Introduction: Provides the research overview, presents the problem statement and outlines the objectives and significance of the study.

Literature Review: Presents an outlook of machine learning SMS classification based on literature and outline the observed research gaps and best practices.

Research Methods: Describes the methods employed to collect data, pre-process it and carry out the research experiments.

Research Findings: Presents the research findings.

Data Analysis and Discussions: Presents an analysis, interpretations and discussions of the research data and findings.

Recommendations and Conclusion: Outlines whether the study objectives were met, presents recommendations based on the analysis and interpretations of research findings then concludes with a brief summary of the entire research.

1.9 Summary

Mobile banking users, locally and globally are suffering huge losses in finance and goods at the hands of EFT SMS scammers. This study set out to train NB, SVM and RF machine learning classifiers to predict ham and EFT scam SMS classes then evaluated their performances in order to determine which of these classifiers can be implemented into effective practical solution to safeguard m-banking users from EFT SMS scams.

The next chapter will focus on scholarly works related to the study subject.

2. LITERATURE REVIEW

This chapter presents an outlook of machine learning SMS classification based on literatures. Predisposed to discover best methods to fulfil the study objectives, the chapter explored and examined similar works, paying special attention to approaches of collecting SMS corpora, identifying and extracting features for classifying SMSes and evaluating machine learning classifier. Furthermore, the chapter captured strengths and shortcomings for the aforementioned methods alongside the observed knowledge gaps and best practices.

2.1 SMS Corpora

Studying SMS classification requires access to an appropriate dataset or corpus of SMSes depending on the study subject. With respect to this study, access to a corpus of ham and EFT scam SMSes is required. Attaining access to a suitable corpus often presents the first challenge that scholars must tackle prior to embarking on SMS classification studies. Abdulhamid (2017) highlighted the difficulties associated with obtaining appropriate corpus for SMS classification, narrating "accessibility to a requisite dataset constitutes one of the challenges researchers often face in successfully carrying out research on filtering or classifying SMS spam messages" (p. 15665).

Literature related to this research were surveyed to identify different approaches normally employed to acquire SMS corpora for machine learning classification. These approaches are categorised into three main groups described in section 2.1.1 through 2.1.3.

2.1.1 Obtaining SMSes from Public Corpora

This approach involves acquiring ham and spam SMS datasets from publicly available corpora compiled by either other researchers or third parties. Several SMS classification studies revealed to have used ham and spam SMS datasets derived from public corpora (Ahmed, Guan, & Chung, 2014; Ezpeleta et al., 2017; Tan, Goharian, & Sherr, 2012).

Some of the renowned examples of publicly available SMS corpora are the *University of California Irvine (UCI)* and the *National University of Singapore (NUS)* corpora (Abdulhamid et al., 2017). The *UCI* corpus comprises of spam SMSes, while the *NUS* corpus is composed of both spam and ham SMSes. All these two corpora were created specifically for research purposes.

Obtaining SMSes from publicly available corpora have an advantage that it is one of the easiest ways to obtain the requisite datasets. This approach however also tends to have some major limitations. Chen and Kan (2012) pointed out that there is a scarcity of appropriate public SMS corpora suiting specific studies. Nuruzzaman, Lee, and Choi (2011) highlighted its other limitation, revealing that public SMS corpora often lacks important details such as SMS sender numbers due to privacy concerns.

Public SMS corpora also tends to be affected by a major shortcoming that they often contain outdated spam SMSes. Using most recent spam SMSes allows developing classifier models based on features from latest spams. This is particularly important because spams are continuously changing as spammers employ new tactics to circumvent existing spam detection and filtering systems. An application developed

based on old spams would likely be less effective in detecting and filtering the latest spams.

In the context of this study, the literature analysis has not identified any existing public corpus of EFT scam SMSes. This made the approach of obtaining SMSes from public corpora not a suitable data collection method in this study.

2.1.2 Extracting SMSes from Public Web-Based Sources

This approach involves acquiring spam SMSes from sources such as websites and web forums where the public report and share spams. In employing this approach, researchers often manually extract spam SMSes from websites such as GrumbleText, a United Kingdom (UK) web forum where users report spams (Almeida et al., 2011; Junaid & Farooq, 2011).

Almeida et al. (2011) pointed out that the main drawback for this SMS collection approach is the difficult and time-consuming process of going through several pages to identify spam SMSes. One of its advantage is that it allows researchers to collect both older and more recent spam SMSes. With respect to this work, this method looked appropriate for collecting EFT scam SMSes.

2.1.3 Collecting SMSes Directly from Users

This approach allows researchers to collect their own corpus of ham and spam SMSes directly from volunteering users. Chen and Kan (2012) carried out a study that used this method to compile a corpus for use by other researchers. Another notable study

that used this method is by Mujtaba and Yasin (2014), who used several mobile networks to collect SMSes from volunteers.

Analysing studies that obtained SMSes directly from users, indicates that this technique is more suitable to use in the instances where there are no appropriate public ham or spam SMSes corpora. This was demonstrated by Shahi and Yadav (2014) who wanted to study SMS spam filtering for Nepali texts but no public SMS corpora for Nepali language existed. When using this approach, specific attention has to be paid to privacy issues to address the contributors' privacy concerns (Chen & Kan, 2012).

In addition to allowing researchers obtain and use latest spam SMSes, another observed advantage for this technique is that it allows researchers access to details related to SMS senders. Although they might be useful in classification, details such as sender numbers are often missing when other methods such obtaining SMSes from public corpora are used during SMS collection.

Yadav, Kumaraguru, Goyal, Gupta, and Naik (2011) highlighted some of the shortcomings associated with this SMS collection method, pointing out that using it comes with the difficulty to obtain a significant number of spam SMSes. Following such realisation, they decided to use incentives in form of food coupons in order to encourage volunteers to contribute SMSes to the spam corpus. In the end, this workaround allowed them to collect over 2000 unique spam SMSes. They pointed out that spam SMS duplicates was another problem they encountered. They explained that nearly 50% of the initial 4000 spam SMSes they received from the users were duplicates, which was because spammers often send the same messages to a large

group of users. This revelation provides an alert to prospective researchers that wants to collect SMSes directly from users to anticipate and plan a work around the possible spam duplicates problem, which represents a good practice.

In attempts to evade shortcomings associated with individual methods of obtaining or collecting the required SMS datasets, some researchers try to use more than one method. Almeida et al. (2011) and Choudhary and Jain (2017) did this by using publically available SMS corpora and also collecting SMSes directly from volunteers. This research sought to emulate this best practice, and utilised the method of extracting SMSes from public web-based sources and collecting SMSes directly from users.

2.2 SMS Feature Representation and Feature Extraction

Cormack (2008) explained that prior to applying machine learning to classify texts; the texts have to be first represented as a collection of features. These features may be derived from the text or from extrinsic information related to the text. During training, the features are used to learn the classifier to predict the text classes. While during evaluation and application, the learnt model is used to predict classes for the test texts or new texts defined using the same feature representation as the text used to train the classifier.

Choosing appropriate features to represent the text is crucial to classification because it directly affects the resultant model's performance. Hidalgo et al. (2006) underscored the aforementioned disclosure, making the following assertion: "A bad representation of the problem data may lead to a classifier of poor quality and accuracy" (p. 107).

Surveying the literature revealed that two general approaches are employed to determine the features to represent SMSes prior to applying machine learning classification. One of the approaches is whereby the researchers or implementers identify and determine which specific features they would use to represent and to classify the SMSes (Choudhary & Jain, 2017). The other approach involves representing and classifying SMSes using features extracted from SMSes following standard methods such as Bag of Words (BoW) and tokenisation, among others (Shahi et al., 2014; Shirani-Mehr, 2012).

To identify features to represent and classify SMSes by oneself requires studying and analysing the features to determine the distinctive ones that sets the different classes apart (Choudhary & Jain, 2017; Nuruzzaman et al., 2011; Yadav et al., 2011). Features deemed to separate different SMS classes are those understood to be suggestive of an SMS being either a candidate of one class or the other. After identifying the features, methods such as binary representation can be used to indicate the presence or absence of the feature in each SMS comprising the dataset.

Following the approach described in the preceding paragraph, Choudhary and Jain (2017) publicised that after an in-depth study of the characteristic of spam messages, they identified features such as the presence of mathematical symbols, URLs, dots, special symbols, emoticons and mobile numbers, among others, as sufficient to classify ham and spam SMSes. Using the aforementioned features to represent their SMS datasets, they were able to achieve 96.5% True Positive Rate (TPR) with a Random Forest (RF) classifier. They defined the features representation using equation 2.1.

$$S_i = \begin{cases} 1 \\ 0 \end{cases}, \qquad (2.1)$$

where S_i is an i_{th} feature, a binary 1 is indicative of the feature presence in an SMS while a binary 0 specifies its absence. The feature vector used in this case comprised of ten features.

In a similar manner, Nuruzzaman et al. (2011) defined features indicative of whether an SMS is a spam or not, using a vector table and binary representation. The vector table constituted of five features 'buy', 'free', 'Viagra', 'SMS' and 'book' as depicted in Table 2.1.

Table 2.1: Features vector table

SMS ID	Type	Word Attributes				
		buy	*free*	*Viagra*	*SMS*	*book*
SMS 1	Spam	1	1	1	0	0
SMS 2	Spam	0	1	0	1	0
SMS 3	Ham	1	0	0	0	1

Yadav et al. (2011) also identified features to classify ham and spam SMSes. They observed that special characters such as '/' are more frequent in spams and that spam SMSes have higher average word length and a higher probability of numeric words. The observation led them to identify twenty features among which the 'count of spam words', 'count of /' and 'average word length' proved more influential on the classifiers' efficacy.

Examining the works where researchers determined which features to use to represent and classify SMSes as described in the preceding paragraphs indicates that there appears to be no guideline or standard rules that they follow. This makes using this approach difficult especially for new researchers or those less acquainted with machine learning text classification features. This method could hence be viewed as more suitable only when a researcher is well accustomed to ham and spam SMS characteristics and machine learning classification features. Nuruzzaman et al. (2011) have highlighted another limitation for this technique, explaining that it is time-consuming.

The second approach involves extracting features from the SMSes following standard methods such as BoW, tokenisation and one-dimensional ternary patterns (1D-TP) then applying classification based on such features (Günal, 2012; Kaya & Ertuğrul, 2016; Shirani-Mehr, 2012). Besides the fact that these methods tend to extract features that allow effective text classification, they often produce a very large number of features, which if all used could result in slower classifiers.

Studying a hybrid feature selection for text classification, Gunal (2012) presented a comprehensive description of the BoW feature representation. The study explained that BoW representations ignore the actual ordering of words or terms in a message and only consider the term occurrences. Each term constitutes an individual feature that can be assigned a weight representing its significance to the class. The most widely used weighting methods with BoW are terms frequency (TF) and term frequency-inverse document frequency (TF-IDF). The TF weighting only considers the number of occurrences of terms in messages, while the TF-IDF scales down the

TF weights by considering the number of messages in the collection that contains the terms. The WEKA '*StringToWordVector*' unsupervised filter converts text string contents to a similar representation as BoW.

The 1D-TP also referred to as local ternary pattern (Tp) converts the message characters to their 8-bits Unicode Transformation Format (UTF-8) values (Kaya & Ertuğrul, 2016). The character UTF-8 values are compared to that of their neighbours' and the result is expressed as a decimal number. The number of neighbours of a character are represented by a P parameter. The $P/2$ characters before and after the central character are assigned as neighbours of that character. The characters UTF-8 values comparisons are done based on equation 2.2.

$$Tp = \begin{cases} 1 & Pc > Pi + \beta \\ 0 & Pc \leq Pi + \beta \ \ and \ Pc \geq Pi - \beta \\ -1 & Pc < Pi - \beta \end{cases}, \qquad (2.2)$$

where Tp is a ternary pattern, Pc represents a central character, Pi represents neighbouring characters P_0, P_1, P_2, P_3 on the left and P_4, P_5, P_6, P_7 on the right of Pc and β is a threshold parameter chosen by the user.

Kaya and Ertuğrul (2016) highlighted one of the drawback for the 1D-TP method, describing that it presents a challenge in determining the optimal P and β parameters. With this drawback in mind and the observation that the 1D-TP feature representations have not been commonly used in similar works, this study opted to not employ it for ham and EFT scam SMS feature representation and classification.

Tokenisation can also be used to establish the features to represent and classify SMSes. Almeida (2011) used two different tokenisation methods. The first method defined a

token as starting with a printable character followed by any number of alphanumeric characters, excluding dots, commas, and colons from the middle of the pattern. This tokenization method allowed domain texts such as URLs to be split at dots so that the classifier can recognize a domain even if the subdomains vary. This tokenization method appears more appropriate for classifying phishing SMSes which normally contains URLs and web addresses. The second tokenization method defined a token as any sequence of characters separated by blanks, tabs, returns, dots, commas, colons and dashes. This tokenization served to preserve other symbols in the message that may help separate spams from ham SMSes. This tokenization looked suitable to represent SMSes that barely contain URLs.

In view of strengths and shortcomings associated with the two identified approaches for representing and classifying SMSes, this study looked to employ both approaches. It attempted to determine features that appear to distinguish ham and EFT scam SMSes based on their observable characteristics and to employ feature extraction to extract important textual features from the SMSes body contents.

2.3 Features for Optimal Classification

Not all features used to represent texts or SMSes contribute meaningful information towards classification. Furthermore, text representation using methods such as BoW often produce a very large number of features, which if all used for classification can adversely impact classification accuracy (CA) and runtime (Shahi et al., 2014; Uysal, Gunal, Ergin, & Sora Gunal, 2013). Hence, it is imperative to analyse the features representing the texts to determine which ones allow the classifiers to perform optimally. Stressing the need to only use features that allow optimal classification,

Suleiman and Al-Naymat (2017) affirmed that features that are less significant to classification only degrades the classifier performance without offering any improvement to the CA.

Features that allow classifier models to perform optimally can be determined in various ways. Gunal (2012) explained that they can be identified either by applying transformation or selection techniques to the features set. Feature transformation techniques project the original feature space or set to a lower-dimensional subspace that carries more relevant or discriminative information. Feature selection, on the other hand, involves analysing and choosing features that are more relevant to classification from the original feature space without changing their values.

Uysal et al. (2013) indicated that besides the existence of various feature selection techniques such as filters, wrappers and embedded methods, text classification appears to predominantly use filtering methods. They argued that filtering can be done independently from the classifiers and it offers computational simplicity compared to other methods. This trait is desirable for classifier models intended to run on mobile devices, which have limited resources. Informed by this key revelation, this study opted to use filtering to select the features that permit optimal ham and EFT scam SMSes classification.

Gunal (2012) noted that filters assess the relevance of the features using scoring schemes such as information gain (IG). Given by equation 2.3, the IG value for a feature A can be defined as a reduction in the entropy that is achieved by learning A (Azhagusundari & Thanamani, 2013).

$$IG(A) = H(S) - \sum_i \frac{S_i}{S} H(S_i), \quad (2.3)$$

where $H(S)$ is the entropy of the given dataset defined by equation 2.4 and $H(S_i)$ is the entropy of the i^{th} subset generated by partitioning S based on feature A.

$$H(S) = -\sum_{i=0}^{n} P(i) log P(i) , \quad (2.4)$$

where $P(i)$ is the probability of class i in the data group S, while n is the number of classes.

Some scholars also reported to have used methods such as text normalization (Aragão et al., 2016) prior to applying feature selection. These methods also contribute to a reduction in the number of less significant features. This study chose to apply text normalization during the feature extraction process since it eliminates the number of unnecessary features created if both lower and upper case characters are used.

2.4 Machine Learning Text Classification Algorithms

There are numerous machine learning algorithms that can be used to classify textual contents, including SMSes. Sections 2.4.1 through 2.4.4 describes and summarises various machine learning algorithms that were considered to train and evaluate the ham and EFT scam SMS classifiers in this study.

2.4.1 Naïve Bayes

The NB is a simple probabilistic classifier based on the Bayes Theorem with strong naïve independence assumptions. These assumptions treat each word as single, independent and mutually exclusive. The NB generally performs classification based on terms frequency or words occurrence.

A study by Nuruzzaman et al. (2011) revealed that Naïve Bayes (NB) is among machine learning algorithms that are broadly used for SMS classification. The work highlighted simplicity as the main advantage of the NB classifiers. Ahmed et al. (2014) cited speed as another advantage for the NB classifiers, revealing that for their experiments, it took an average of 0.13s to classify an SMS while the then state of art NB text classifiers took around 70μs. The simplicity and speed traits for NB are particularly desirable for classifiers intended for mobile devices.

The main limitation of NB classifier is associated with its assumption about the features' independence, which is not always the actual case (Hedieh, Parast, & Akbari, 2016).

2.4.2 Support Vector Machine

The support vector machine (SVM) is a non-probabilistic binary algorithm that can be used for classification and regression analysis (Hedieh et al., 2016). Similar to the NB, the SVM is also reported to be broadly used in SMS classification (Reaves, Blue, Tian, Traynor, & Butler, 2016; Shahi et al., 2014).

Rohith (2018) explained that SVM employs supervised learning and uses a set of class-labelled training data points to build a model to predict classes for new data points. The model is built by representing the class-labelled data as points in space, and subsequently mapping them such that points from different classes are divided by a clear gap that is as wide as possible. The new data points are then mapped into the same space, and their classes predicted based on which side of the gap they fall.

Several works on SMS spam detection that used SVM, reported CA above 90% (Reaves et al., 2016; Shahi et al., 2014). When compared to other classifiers such as NB, the SVM often requires a long training time (Hedieh et al., 2016). This is one of the main limitation for the SVM classifiers.

2.4.3 K-Nearest Neighbours

Uysal et al. (2013) explained that for each new data instance requiring classification, the k-nearest neighbours (KNN) classifier determines the closest training example and classifies the new instance based on that example. The new data instance being classified is allocated a class that is most common for its k-nearest neighbours. For instance, if k = 4, and three of the four closest neighbours to the SMS being classified belongs to ham class, then that SMS would be classified as a ham SMS. For special cases where k = 1, meaning only one closest neighbour is being considered, the new data item is simply assigned to the class for its only nearest neighbour.

A study by Hedieh et al. (2016) highlighted that while the KNN classifiers are easy to implement and update, their main limitation is that their performance tends to degrade with increase in noise in the training data. The KNN did not appear to have been regularly used for SMS classification.

2.4.4 Random Forest

Random Forest (RF) also appear to be frequently used for SMS classification, with several scholars underlining its efficiency in classifying ham and spam SMSes (Shirani-Mehr, 2012; Suleiman & Al-Naymat, 2017). Choudhary and Jain (2017)

explained that RF classifiers construct a forest of decision trees during the training phase and makes it random. Tekerek (2018) further expounded that when splitting a node, the RF does not look for the most important feature, but searches for the best feature among a random subset of features, which diversifies the model and enables it to perform better. The main limitation of RF is that it tends to be slow when the number of trees is large (Donges, 2019).

From the four discussed algorithms; NB, SVM and RF classifiers were trained and evaluated in this research. The choice followed from the three classifiers reported common use in similar works and their highlighted strengths as discussed and explained from section 2.4.1 to 2.4.4. The research experiments were done using WEKA platform, which contains all the selected three algorithms (Bouckaert et al., 2002).

2.5 Classifier Evaluation Metrics

Hossin and Sulaiman (2015) elucidated that selecting suitable metrics to evaluate the performances is central to analysing and comparing various classifiers. Using appropriate evaluation metrics allows determining the best-suited classifiers for a specific problem or application.

Mishra (2018) noted that a classifier models may exhibit a satisfying performance when evaluated against one metric but perform poorly with respect to another. This prompt for the use of several metrics when evaluating classifiers to ensure that their performances are scrutinised and analysed from a number of perspectives. Classifier performances are commonly evaluated based on confusion matrix fundamental metrics

of True Positives (TP), True Negatives (TN), False positives (FP) and False negatives (FN) (Abdulhamid et al., 2017; Hedieh et al., 2016; Mahmoud & Mahfouz, 2012). Table 2.2 depicts a confusion matrix and its four basic metrics used to evaluate classifiers (George, 2019).

Table 2.2: Confusion matrix

		PREDICTED CLASS	
		EFT Scam	**Ham**
ACTUAL CLASS	**EFT Scam**	TP	FN
	Ham	FP	TN

Mahmoud and Mahfouz (2012) described the confusion matrix`s metrics with respect to ham and general spam SMSes as follows:

TP: The number of spam SMSes that correctly classified.

TN: The number of ham SMSes that are correctly classified.

FP: The number of ham SMSes that are falsely classified.

FN: The number of spam SMSes that are falsely classified.

Other common metrics central to evaluating classifier performances can be derived and computed from the four basic metrics of the confusion matrix. Suleiman and Al-Naymat (2017) and Aragão et al. (2016) defined these metrics and gave their respective equations as follows:

FP Rate (FPR): refers to the rate of ham SMSes misclassifications.

$$FPR = \frac{FP}{FP+TN} \quad (2.5)$$

FN Rate (FNR): refers to the rate of spam SMSes misclassifications.

$$FNR = \frac{FN}{FN+TP} \quad (2.6)$$

Classification accuracy (CA): gives the ratio of the number of correctly classified SMSes to the total number of input SMSes to the classifier.

$$CA = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.7)$$

Precision: denotes the ratio of messages classified as spams that are actually spams. Precision shows the exact correctness.

$$Precision = \frac{TP}{TP+FP} \quad (2.8)$$

Recall: denotes the ratio of actual spams that are correctly classified. Recall shows the completeness.

$$Recall = \frac{TP}{TP+FN} \quad (2.9)$$

F1-measure: represents the harmonic mean of Precision and Recall.

$$F1 - measuse = \frac{2 \text{ x } Precision \text{ x } Recall}{Precision+Recall} \quad (2.10)$$

2.6 Binary and Multiclass classification

As explained in the Background of the Study section, SMSes falls under two general classes, the ham and the spams. The ham class includes all normal, desired and legitimate SMSes while the spams class includes advertisements, promotions, scams (including EFT scams) and other unsolicited SMSes. For practical reasons SMS pam

filtering and detection applications largely employ binary classifications as opposed to multiclass classification (Abdulhamid et al., 2017; Akinyelu & Adewumi, 2014; Ezpeleta, Garitano, Zurutuza, & Hidalgo, 2017). This is because multiclass classifier models generally utilise more processing and storage resources (which are more limited for mobile devices) compared to binary classifier models.

2.7 Summary

Three main approaches; obtaining SMSes from public corpora, extracting SMSes from public web-based sources and collecting SMSes directly from users are used to obtain requisite corpora for studying ham and spam SMS classification. No suitable public corpora for ham and EFT scam SMEs were identified, hence this study could only collect SMSes either from public web-based sources or directly from users. Literature indicate that NB, SVM and RF classifiers are commonly used to classify ham and general spam SMSes. The study looked to use FPR, FNR, CA, Precision, Recall and F1-measure performance metrics in conjunction with the confusion matrix of TP, TN, FN, and FP to evaluate the classifiers performance.

The next chapter focuses on research methods.

3. RESEARCH METHODS

This chapter outlines the methods employed to collect and pre-process the SMS corpora. The chapter further explains and elaborates how the research experiments were carried out. First, the used research design is explained, then the target population is defined before a detailed presentation of the experimental procedures is given.

3.1 Research Design

The study employed a quantitative experimental research design. Ham and EFT scam SMSes were collected and WEKA (Waikato Environment for Knowledge Analysis) data mining software platform was used to perform the experiments to extract, analyse and determine machine learning features that can be used to optimally classify the SMSes. As outlined in the Literature Review chapter NB, SVM and RF were selected from the four classifiers that were considered because of their common uses in SMS classification and their reported strengths. After training to predict the ham and EFT scam SMS classes, the three classifier model performances were evaluated and quantitatively analysed based on TP, TN, FP, FN, CA and precision among other metrics. The model training and evaluation followed a 10-fold cross-validation (CV) approach.

3.2 Population

The study population comprised of FNB Namibia m-banking service users. As alluded to in the Problem Statement section, most of the local EFT scam reports involve FNB, making their m-banking service users the population of interest to the research. This

population is virtually made up of all FNB Namibia bank account holders and all individuals with valid Namibian SIM numbers. The latter is due to the fact that an individual only requires a valid SIM number to have access to EFT services such as e-wallet, hence to be considered as a m-banking service user.

3.3 Sample

Two sampling techniques, purposive and self-selection sampling were utilized to simplify the data collection process. The purposive sampling technique was used to collect EFT scam SMSes from user posts on Facebook public groups. This technique was further employed to target potential volunteers and users that had come across EFT SMS scams to solicit for their contribution to the associated corpus. Self-selection sampling, on the other hand, was used to grant m-banking users a choice to contribute to the ham corpus. The elaborate steps on how these sampling techniques were utilized are outlined in the Procedures section. A sample consisting of 240 unique ham and EFT scam SMSes was collected. This entire sample of SMSes was used in the study experiments.

3.4 Procedures

The different methods and processes followed to collect and pre-process the ham and EFT scam SMSes and to perform the study experiments are outlined and explained in section 3.4.1 to 3.4.6.

3.4.1 SMS Collection and Raw SMS Feature Representation

Based on the deductions explained in the SMS Corpora section, two approaches were identified as best suited to collect the ham and EFT scam SMSes. The two approaches were implemented and applied as outlined in the following sections 3.4.1.1 and 3.4.1.2.

3.4.1.1 Extracting SMSes from Facebook Public Groups

This method was chiefly employed to gather the EFT scam SMSes and involved searching and extracting them from user posts on Facebook public groups. Mobile banking users have a habit of sharing screenshots of EFT scam SMS contents and scammer mobile numbers on social media platforms such as Facebook to warn other users. Facebook public groups have large user communities, making them an ideal platform to collect EFT scam SMSes. At the time of data collection, the searched Facebook groups have user communities ranging from 10 000 to over 170 000 users. The groups listed in Table 3.1 were searched for EFT scam SMSes.

Table 3.1: Facebook groups searched for EFT scam SMSes

Group Names	
Crime in the City NAMIBIA	Buy or Sell Namibia Central
Namibia Online Advertising	The Market Place Namibia
Unam sample`s	Buy or sell in WINDHOEK $$$$
NSFAF BENEFICIARIES	Buy or Sell Anything Namibia
STREET MARKET WHK	WINDHOEK OPEN MARKET
Buy and Sell Namibia	WINDHOEKBUY AND SELL
Namibia Online Advertising	BUY | TRADE | SELL NAMIBIA
Buy or Get it Sold	Namibia Second Hand Sales
Namibian Classifieds by Namlist	Buy or Sell Namibia North

The flow chart and accompanying explanations in Figure 3.1 describe the explicit steps followed to collect EFT scam SMSes from the groups listed in Table 3.1.

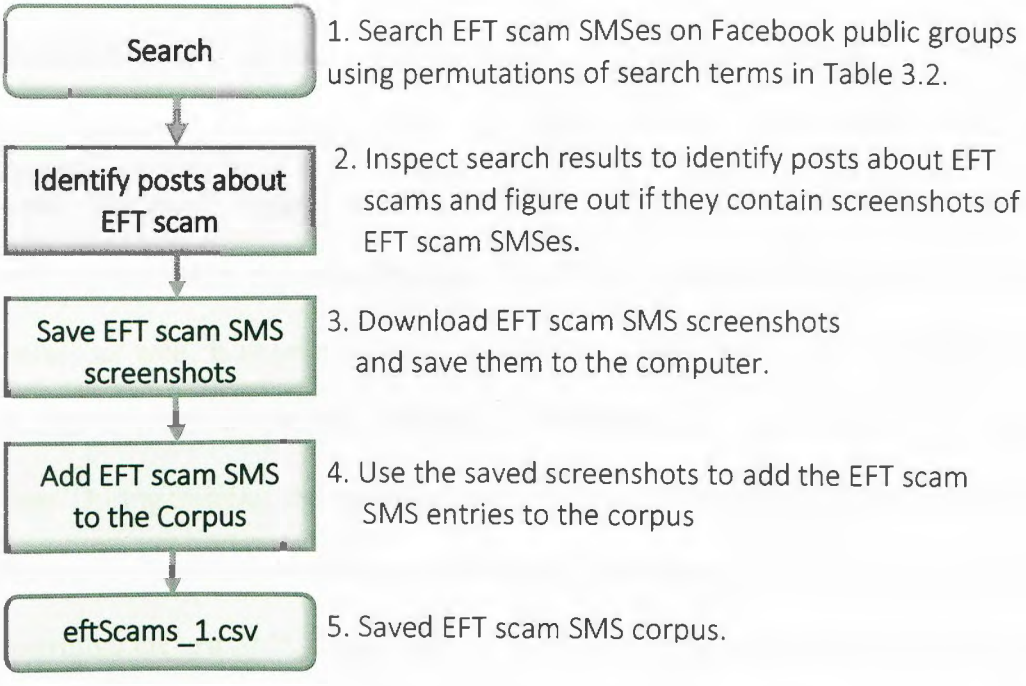


Figure 3.1: Steps followed to collect EFT scam SMSes from Facebook groups

Various permutation of terms listed in Table 3.2 were used to search EFT scam SMSes from user posts on Facebook public groups.

Table 3.2: Used EFT scam SMS search terms

Search Terms	Examples of search terms permutations
FNB	FNB + EFT + deposit + SMS + notification + scams
EFT/ e-wallet/ ewallet	e-wallet + SMS + scams
Deposit	ewallet + notification + frauds
SMS	ewallet + deposit + scam
Notification	FNB + e-wallet + frauds
scam(s)/ fraud(s)/thief	ewallet + thief

As revealed in the Literature Review chapter, choosing an appropriate text representation is paramount to machine learning classification. The features used in this study to represent the raw or unprocessed SMSes were identified following the findings from extensive examination of ham and EFT scam SMS characteristics and a consideration of features used in similar works. The used features are 'senderNumLen', 'senderSavedAs+362626', 'content' and 'contentLen' and were used together with the class attribute 'smsClass'. Although 'smsClass' is literally a feature as well, it is referred to as an attribute in this work to set it apart from the normal features. It specifies whether an SMS belongs to the ham or the EFT scam class. During training, the values for the 'smsClass' allow the classifiers to learn which features are associated with each SMS class, while during evaluation they are used to determine the test SMS classes that are predicted correctly and those that are not. The four features and the class attribute used to represent raw SMSes are described in Table 3.3.

Table 3.3: Description of features used to represent raw SMSes

Features or Attribute	**Description**
senderNumLen	Represents the total number of digits that constitutes the mobile number of the SMS source.
senderSavedAs+362626	Denotes whether the SMS sender number is saved on the victim's handset using +362626 as a name.
content	Contains the string content of the SMS body.
contentLen	Represents the number of characters in the SMS body.
smsClass	Defines whether the SMS belongs to Ham or EFT scam class.

3.4.1.2 Collecting SMSes from volunteers

This approach involved asking volunteers to contribute SMSes to the research and was used to collect both ham and EFT scam SMSes. Ham SMSes normally contain private user contents, which can raise privacy concerns as alluded to in the SMS Corpora section of the Literature Review. To address such concerns, the SMS collection from volunteers was done with their duly agreement and consent. Additionally, all volunteers were informed of the research intents, and that dedicated efforts would be made to ensure their anonymity and to protect the collected SMSes data. The use of this approach to collect the EFT scam SMSes was mainly motivated by the need to complement the method of gathering them from the Facebook public groups. This was to ensure that a reasonably sized dataset of EFT scam SMSes was obtained.

In the application of this SMS collection approach, firstly a Google form outlining the study details and giving instructions on how volunteers can contribute SMSes was created alongside a mini-survey. Although the mini-survey did not serve to contribute to the central objectives of the study, it solicited views about EFT SMS scams from the m-banking users' perspectives. Such views aided with the understanding of the research problem and helped with the problem descriptions presented in the introductory chapter.

Electronic invitations to contribute SMSes to the study corpus were compiled and send out to potential volunteers. These invitations, which had embedded URL links to the study details Google form and the mini-survey, were posted on the same Facebook public groups used in the previous section to search for EFT scam SMSes. In addition, they were also posted on WhatsApp groups such as *'Buy and Sell (Windhoek)'*, *'NAM*

SALES & MARKETING' and *'NAMIBIA BUY AND SELL'*. Besides having fewer followers (i.e. below 250 per group), WhatsApp groups provided an alternative route to solicit SMSes from more users, especially those not members of the searched Facebook groups.

The invitation posts and URL links were set as sharable, allowing users to share them using Facebook, WhatsApp, Email or mobile phone SMSes. This permitted reaching out to a large pool of potential volunteers. Figure 3.2 depicts a screenshot with an invitation that was posted on a Facebook public group *'Unam samples'*.

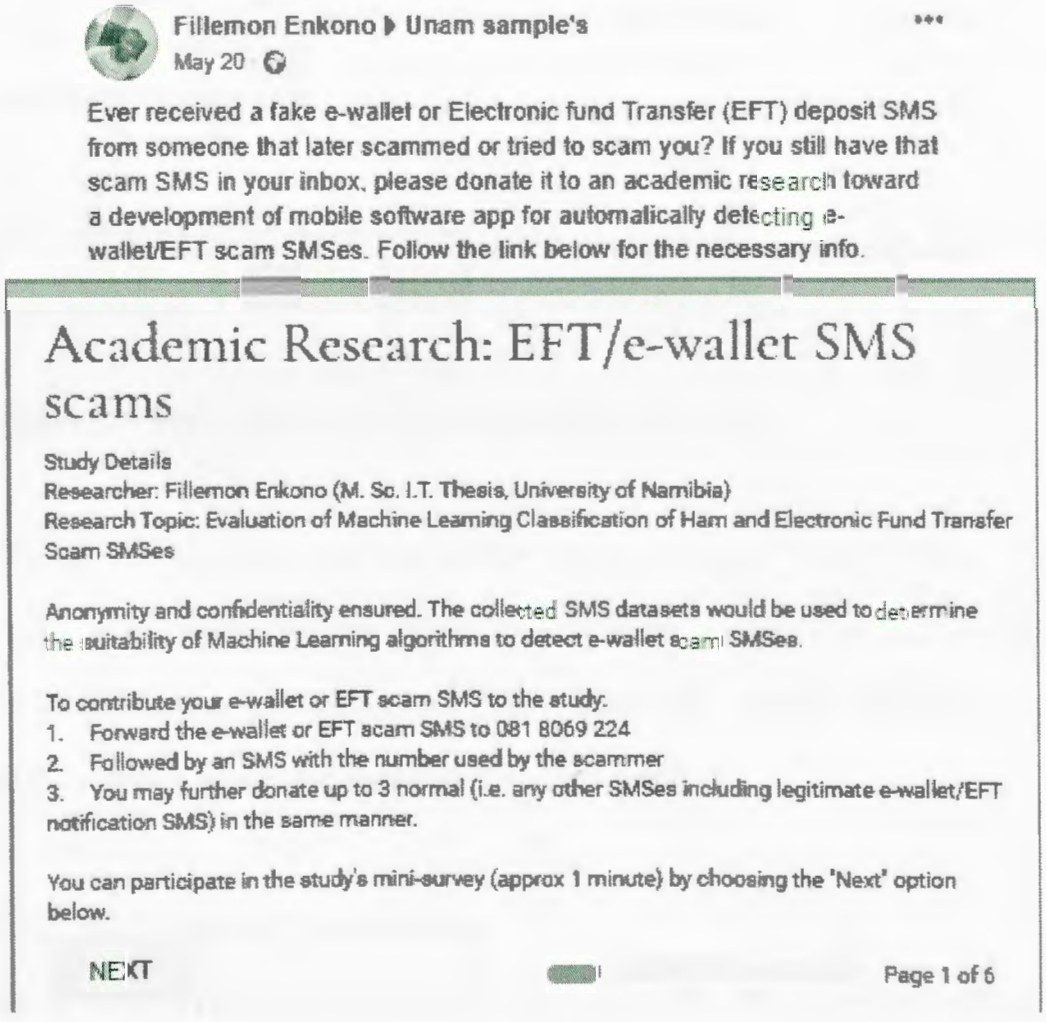


Figure 3.2: **Invitation for volunteers to contribute SMSes**

A larger snapshot of the study details Google form is presented in appendix A.

Figure 3.3 describes the exact steps followed to collect SMSes from volunteers and to compile them into a ham and EFT SMS corpora.

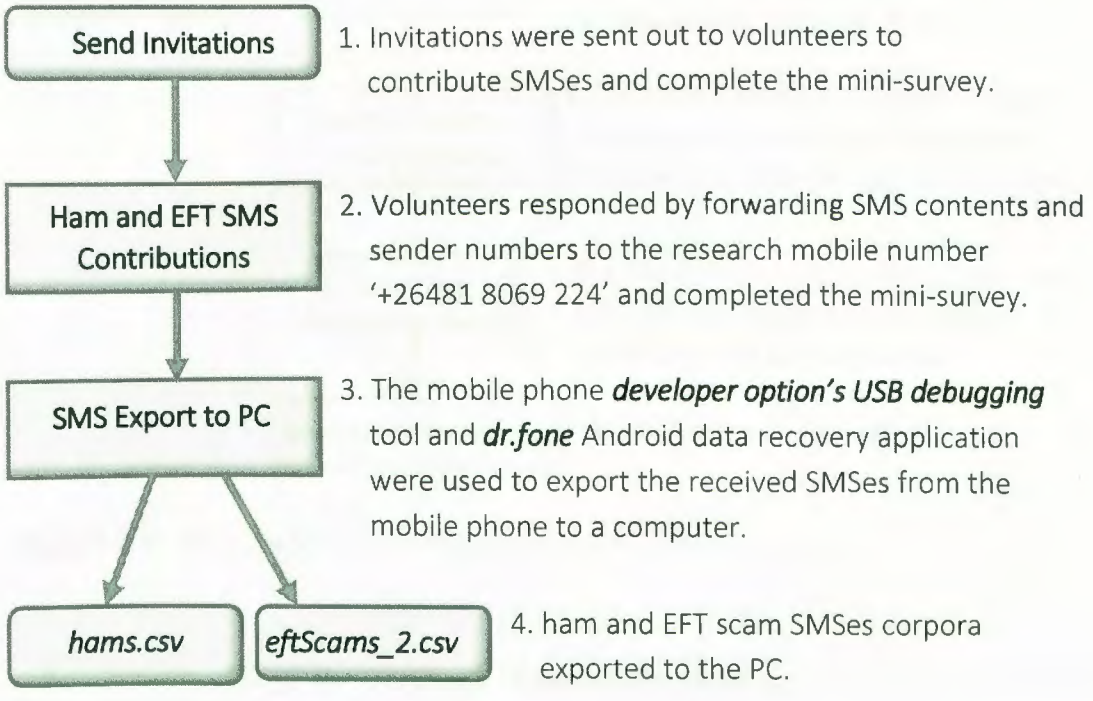


Figure 3.3: Steps followed to collect SMSes from volunteers

The corpora exported from the mobile phone; 'hams.csv' and 'eftScams_2.csv' contained extra SMS features such as, *'time'*, *'smsState'*, *'smsType' and 'threadID'*, in addition to the *'senderNumber'*, *'senderName'* and *'content'* features. The next section elaborates how the extra features where handled.

3.4.2 SMS Corpora Pre-processing

The diagram in Figure 3.4 summarises the pre-processing enacted on the SMS corpora using 'LibreOffice Calc' application.

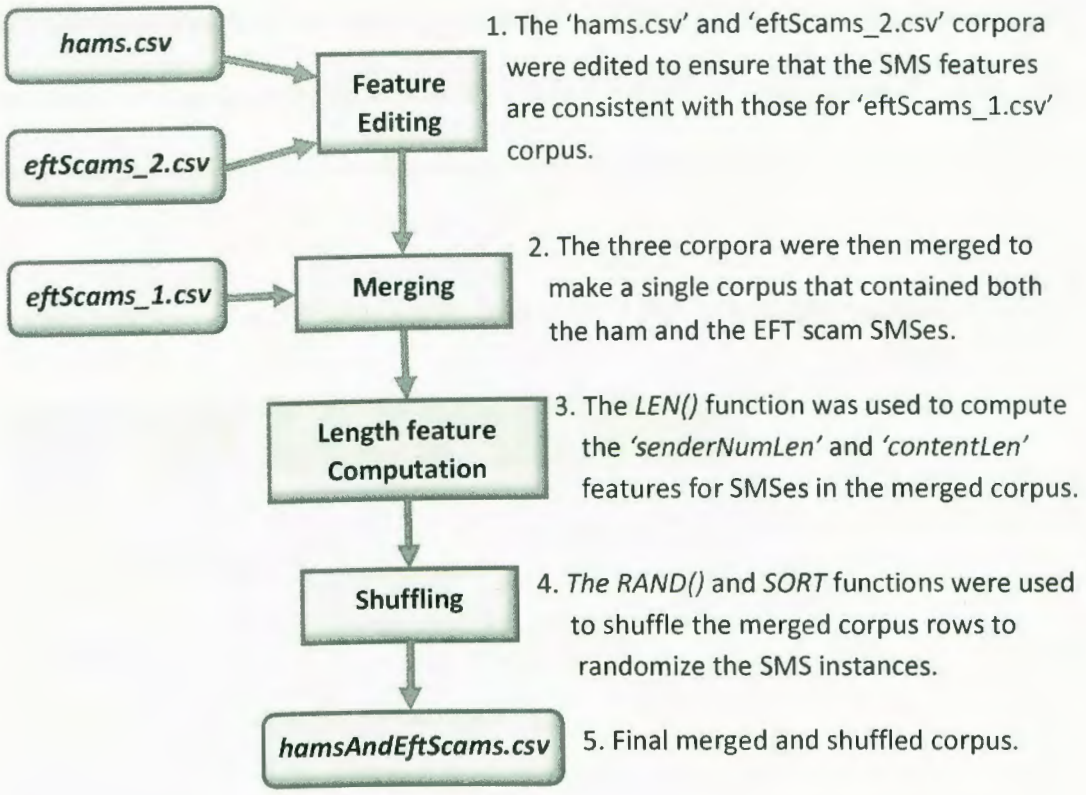


Figure 3.4: SMS corpora pre-processing with 'LibreOffice Calc'

After pre-processing, the resultant 'hamsAndEftScams.csv' corpus which had a *Comma Separated Value (CSV)* data format was converted to 'hamsAndEftScams.arff' that has an *Attribute Relation File Format (ARFF).* The conversion was necessary because unlike *CSV, the ARFF* file format allows explicit definition of feature datatypes to ensure compatibility with machine learning classifiers. The conversions were done using 'Notepad++' text editor and 'WEKA *ARFF'* editor.

The four features and the class attribute defining SMSes in the 'hamsAndEftScams.arff' used three datatypes, '*numeric*', '*nominal*' and '*string*'. WEKA documentation provide definitions for these datatypes as follows (Bouckaert et al., 2002). The features with *numeric* datatype takes real or integer values. The

nominal features take nominal-specifications, listing all the possible feature values while the *string* features take arbitrary textual values. The data type for each of the four features and the class attribute are listed in Table 3.4.

Table 3.4: 'hamsAndEftScams.arff' corpus SMS features or attribute data types

Feature/attribute	Data type
senderNumLen	Numeric
senderSavedAs362626	Nominal
content	String
contentLen	Numeric
smsClass	Nominal

3.4.3 Feature Extraction

WEKA '*StringToWordVector*' unsupervised filter was utilised to extract classification features from the SMSes. The '*StringToWordVector*' filter converts text contents to features similar to that from BoW features. The features were extracted from texts in the bodies of raw SMSes (i.e. SMSes '*content*') and involved converting the texts to words and terms attributes and their number of occurrences in each SMS. Figure 3.5 and the associated descriptions summarises the steps followed to extract the features.

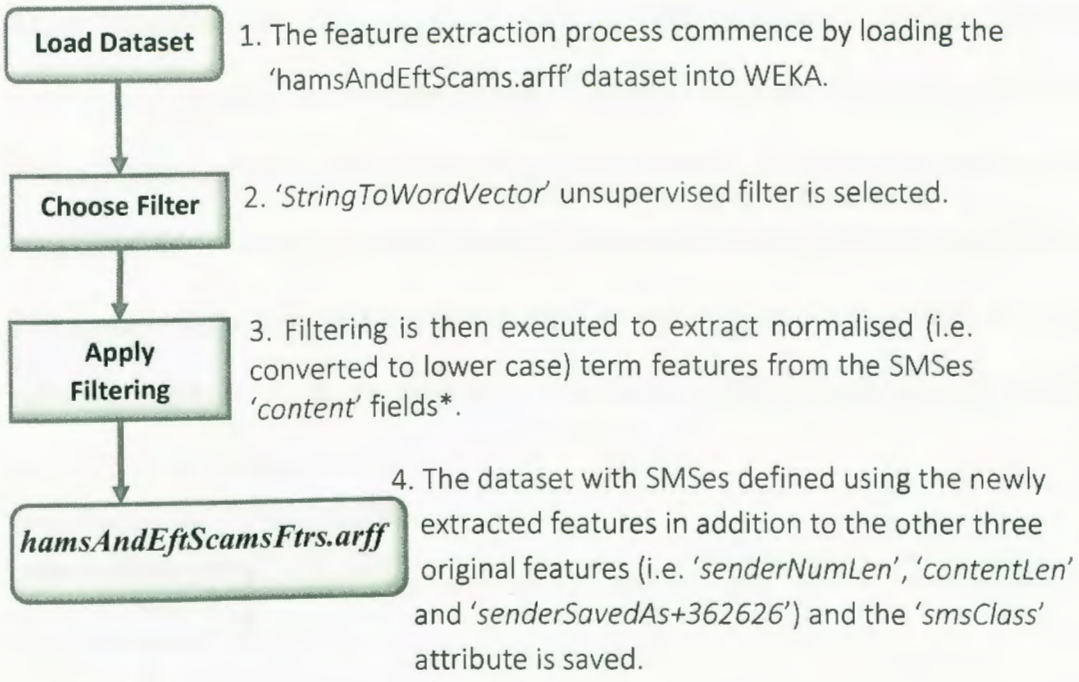


Figure 3.5: Feature extraction process

*When executed the '*StringToWordVector*' filter generates an indexed list of all normalised unique terms and words in the '*content*' fields of SMSes comprising the dataset. Each SMS content is then defined using pairs of indices (that represent the terms and words it contains) and the number of the respective terms occurrences.

3.4.4 Determining Features for Optimal Classification

SMSes in the 'hamsAndEftScamsFtrs.arff' dataset produced from feature extraction were defined using 1226 features and the '*smsClass*' attribute. As acknowledged in the Features for Optimal classification section, not all features contribute meaningful information towards classification. This makes the process of determining features that allow optimal classification essential.

The Information Gain (*IG*) feature selection algorithm was used to evaluate the 1226 features and compute the information they contribute towards the SMS classification. This allowed the features that are less significant toward classifying the SMSes to be eliminated from the set used to define the SMSes prior to classifier training and testing. The graphic depictions and associated descriptions in Figure 3.6 describe the steps followed to determine the features that could allow optimal classification of the ham and EFT scam SMSes.

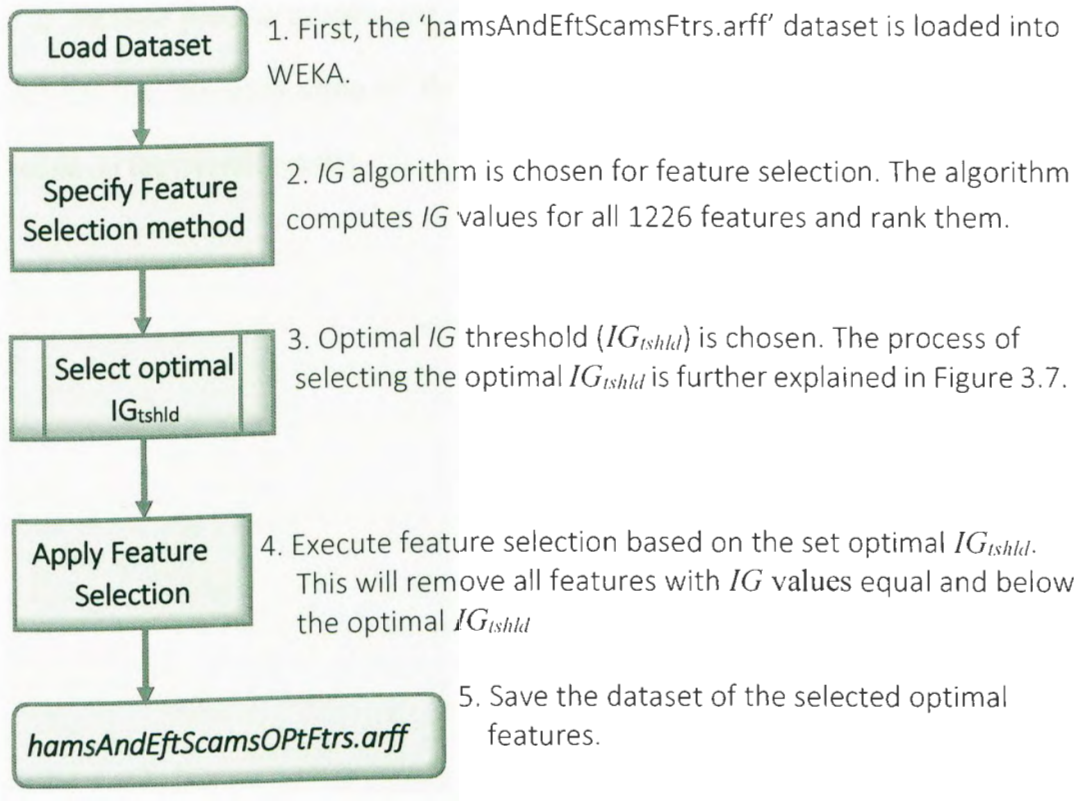


Figure 3.6: Determining features that allow optimal SMS classification

Selecting the IG_{tshld} value that allows optimal classification for a specific problem required employing feature selection with various IG_{tshld} values, followed by training the classifiers and then testing them using SMS datasets defined with the selected features. The CA was used as the evaluation metric when determining the optimal

classification features. The IG values are bound by the [0, 1] set. If all features in the dataset are defined as F and the IG value for an i^{th} feature f_i is denoted as IG_i, then applying IG feature selection with:

a. $IG_{tshld} = x$, *where* $x<0$ will result in a selection of all features F comprising the dataset (i.e. f_i with $1 \geq IG_i \geq 0$)

b. $IG_{tshld} = x$, *where* $0 \leq x \leq 1$ will result in a selection of features f_i with $1 \geq IG_i > IG_{tshld}$.

The detailed process of selecting the optimal IG_{tshld} is described in Figure 3.7. The process was repeated using all three classifiers and the optimal IG_{tshld} was selected based on the overall results.

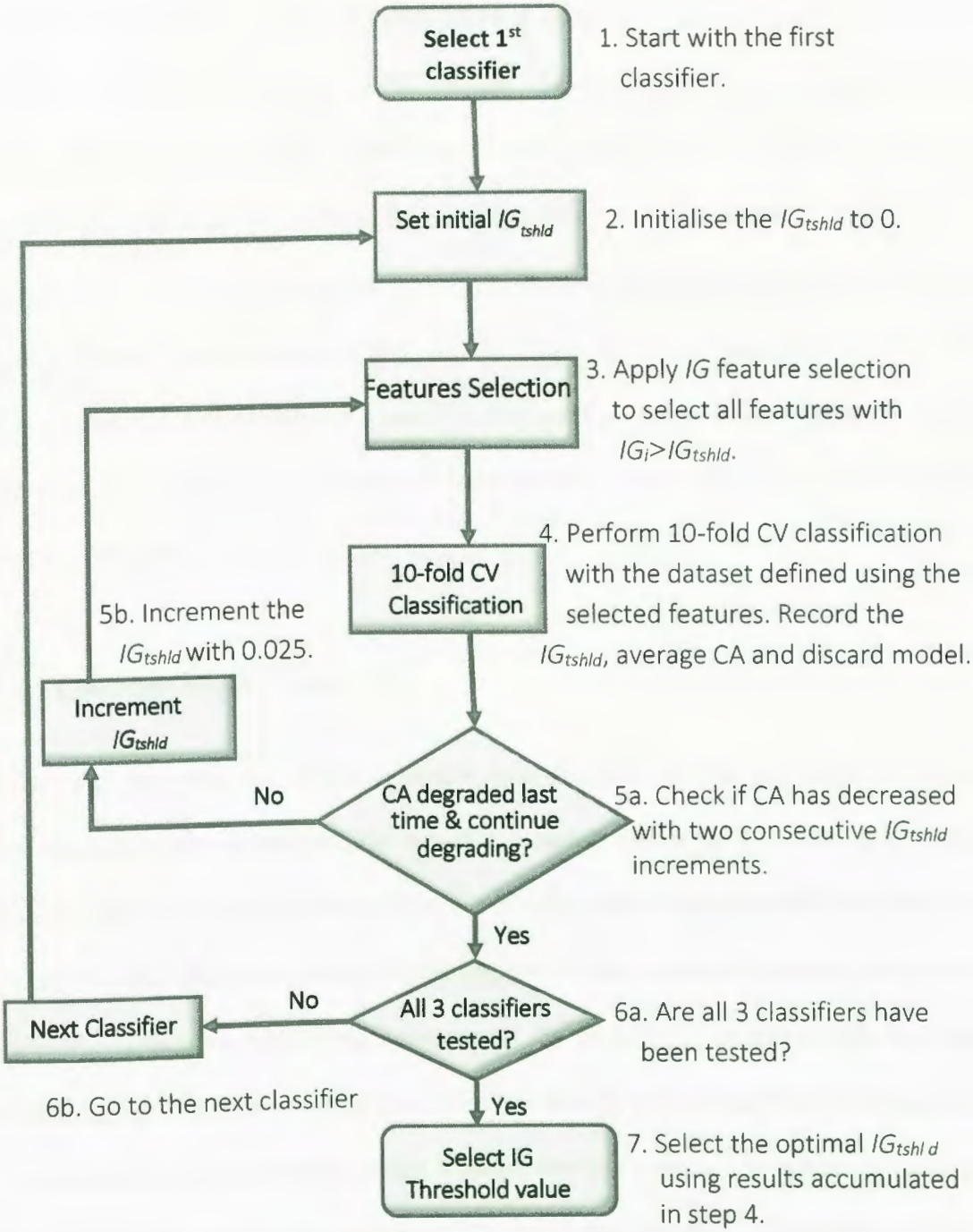


Figure 3.7: Selecting optimal IG_{tshld} value

3.4.5 Classifier Models Pre-evaluation

Pre-evaluating the NB, SVM and RF classifiers followed after the optimal IG_{tshld} value was determined. The pre-evaluation involved validating that the features determined as optimal for classifying the SMSes really allowed the classifiers to perform optimally

compared to when all features are used. This was done by comparing the performances of the models trained using SMS datasets defined using all the features (i.e. *'hamsAndEftScamsFtrs.arff'*) with those trained using datasets defined using the features selected with the optimal IG_{tshld} (i.e. *'hamsAndEftScamsOptFtrs.arff'*). As a measure to save time for the actual evaluation, the classifier training and testing during pre-evaluations used automated 10-fold CV, which performed auto dataset splitting into training and testing folds and used the average CA as a performance metric. The pre-evaluation outcomes are captured with the rest of the results in the Research Findings chapter.

3.4.6 Classifier Models Evaluation

After analysing the pre-valuation results and ascertaining the necessity of using optimal classification features, the actual evaluation followed. The evaluation used datasets defined using the optimal features instead of the entire set of 1226 features. During the classifiers evaluation, it was imperative to inspect and analyse the results for all the 10 training and testing iterations for the 10-fold CV instead of only looking at the average results as was the case with the models pre evaluation and during the experiments to determine the optimal IGtshld. For this reason, a step by step manual 10-fold CV approach was employed. This allowed observing, documenting and analysing the models' performance during each training-testing iteration. Figure 3.8 and the subsequent expansions of points **a.** through **d.** gives an elaborate description of the evaluation process for each classifier.

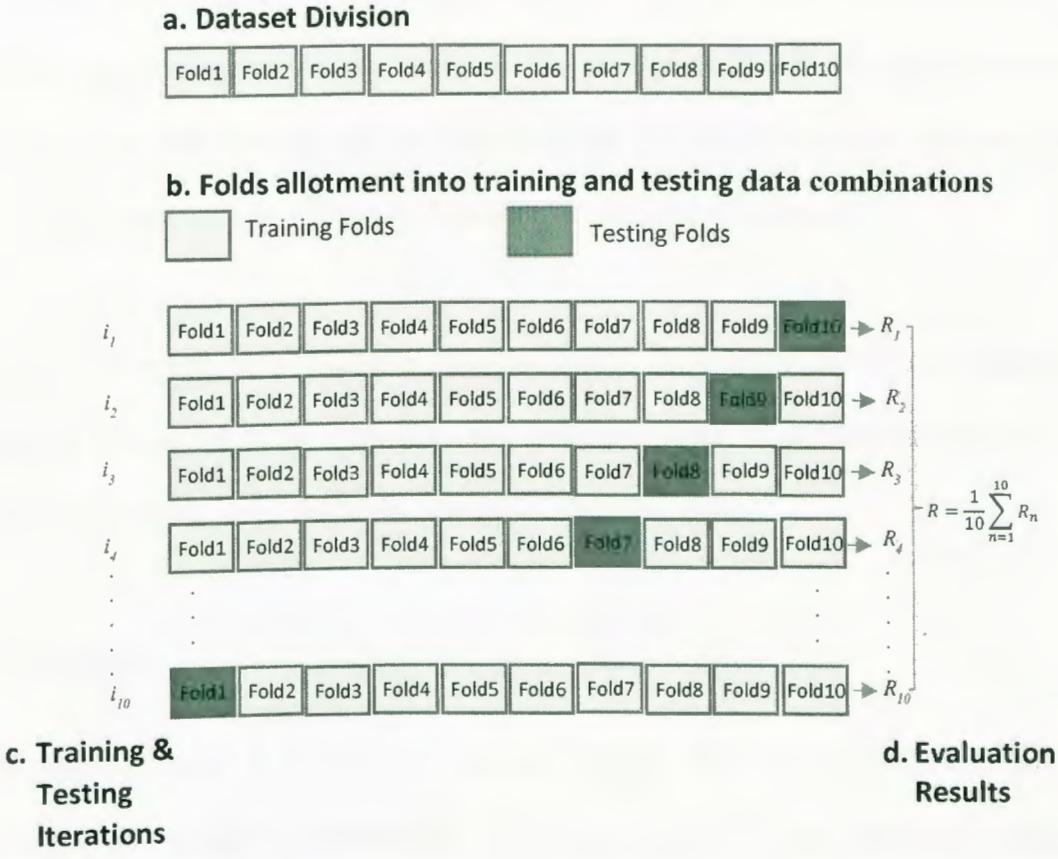


Figure 3.8: Classifier evaluation process

Explanation of points in Figure 3.8:

a. First, the shuffled 'hamsAndEftScamsOptFtrs.arff' dataset was partitioned into 10 equal folds with 24 SMS instances each.

b. The folds were systematically allotted into ten different combinations such that each combination constitute nine folds of training data and one fold of testing data.

c. A series of ten training and testing iterations were then used to learn (or train) each of the three classifiers to predict the ham and EFT scam SMS classes and evaluate the resultant models' performance. At the end of the 10 iterations, each fold has been used nine times in training data and once in testing data to evaluate the model.

d. The evaluation results' confusion matrices of TP, FP, TN and FN were recorded for each training and testing iteration. The average evaluation results were then calculated and subsequently used to compute the values for other evaluations metrics such as FPR, FNR, CA, Precision, Recall and F1-measure.

The evaluation results for the three classifier models are captured in the next section. Upon the conclusion of the evaluation, the three models for classifying ham and EFT scam SMSes were saved for future testing, validations or use.

3.5 Summary

The research used a quantitative research design and utilised purposive and self-selection sampling techniques. The raw SMSes were defined using '*senderNumLen*', '*senderSavedAs362626*', '*content*' and '*contentLen*' features in addition to the '*smsClass*' attribute. Following this, the '*StringToWordVector*' unsupervised filter was utilised to extract words and terms features from the '*content*' texts of SMSes constituting the corpus. Feature selection using IG algorithm was then employed to select features that allow optimal classification. Upon validating the necessity for using optimal classification features, the NB, SVM and RF classifiers were trained then evaluated following a 10-fold cross-validation approach.

The next chapter looks at the findings and results obtained following the methods and procedures that were presented in this chapter.

4. RESEARCH FINDINGS

This chapter presents the findings of the research. It lists and presents the results for the SMS collection and pre-processing, feature extraction and optimisation, and classifiers training and evaluation.

4.1 Collected SMSes

The data collection resulted in a gathering and compilation of a corpus with a total of 240 unique ham and EFT scam SMSes. Figure 4.1 summarises the composition of the collected corpus in terms of ham and EFT scam SMSes.

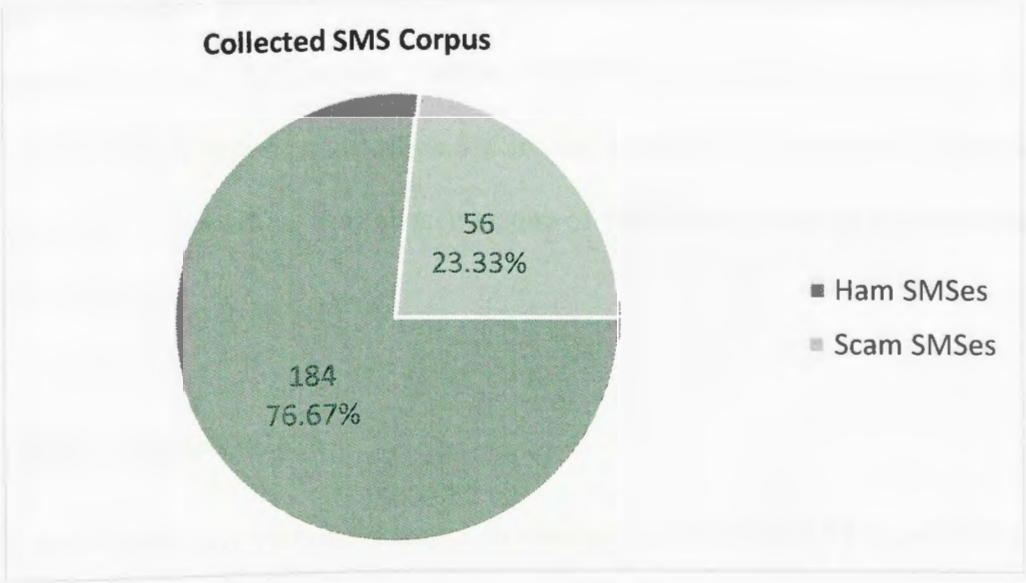


Figure 4.1: Composition of the collected SMS corpus

As Figure 4.1 outlines, the collected corpus comprised of 184 ham and 56 EFT scam SMSes, constituting 76.67% and 23.33% respectively.

4.2 Raw SMSes Representation

The raw SMSes were first represented using appropriate features prior to applying machine learning classification. The choice of each of the features and the class attribute used to represent raw SMSes were identified following these findings:

senderNumLen

Banks use short SMS codes (for example FNB Namibia uses +362626) to send EFT deposit notification SMSes to m-banking users. The short SMS codes have fewer digits compared to the normal mobile SIM numbers (for example +264811234567) which spammers often use to send the EFT scam SMSes. This makes the number of digits comprising the SMS sender number essential to making distinctions between legitimate EFT deposit notification SMSes and the EFT scam SMSes. For this reason, the '*senderNumberLen*' was chosen as one of the features used in the representation of raw SMSes.

senderSavedAs+362626

Spammers employ various trickeries in attempts to make the EFT scam SMSes look as close as possible to the legitimate EFT deposit notification SMSes sent by banks. One of such tricks happens when the EFT SMS scammers attempt to swindle goods from salespersons. Prior to sending the EFT scam SMS to the salesperson, the scammer first requests to use the salesperson's mobile phone and saves their number using +362626 as a name without being noticed. This way when the victim mobile phone receives the scam SMS, it will appear as if it was sent by +362626, the short SMS code used by FNB to acknowledge e-wallet and other EFT transactions. This can

easily convince a salesperson to think that the EFT SMS they receive acknowledge a legitimate payment, causing them to give their products to the scammer. The Facebook post shown in the screenshot in Figure 4.2 gives an example of such an incident.

Figure 4.2: EFT scam SMS sender number saved as +362626 on a victim phone

While some mobile phones display SMS sender numbers alongside the saved sender name on top of each message as the case in Figure 4.2, other phones especially non-smartphones only display the saved SMS sender's name. That makes it impossible to tell whether an SMS was sent by +362626 or by a normal SIM number such as +264811234567 that is saved on the phone using +362626 as a contact name. To account for this scenario, a binary feature '*senderSavedAs+362626*' was included among the features used in the representation of raw SMSes. For an SMS, a binary 1 for this feature indicates that the sender number for that specific SMS is saved on the victim's phone using +362626 as a name while a binary 0 indicates the contrary.

Content

The '*content*' comprises of the SMS body and is a key feature because it contains most of the SMS information. The employed feature extraction operated on the strings contained in the SMS content, capturing all words and terms that could contribute toward classifying the SMSes. Examining the ham SMS contents shows that they do not seem to follow any specific structure; they contain slangs, acronyms and often a mixture of words in different languages. This can be confirmed by examining the words and terms features extracted from the SMS contents in Appendix C. The contents for EFT scam SMSes on the other hand closely follow the structures of legitimate EFT deposit notification SMSes. With exceptions of cases when scammers have made errors when forging the scam SMSes, in most cases the only discernible differences between legitimate and EFT scam SMSes is the sender number and the number of digits constituting the sender number.

ContentLen

A close inspection of the collected SMSes showed that EFT scam SMSes generally have more characters compared to ham SMSes. The '*contentLen*' feature was used to capture such information in the raw SMS representation.

SMSclass

The class attribute labels are used for training the classifiers using supervised learning and also during the model evaluation to validate which instances of the test SMSes have been correctly or wrongly classified. The SMS features may be altered during feature extraction or eliminated during the selection of features for optimal

classification if they do not meet the specified IG_{tshld}. Having been specified as the class attribute in the classifier settings, the values for the '*SMSclass*' attribute are left unaltered by feature extraction and are retained during feature selection.

Table 4.1 depicts the first five SMS instances in the shuffled 'hamsAndEftScams.csv' corpus to illustrate the raw SMSes representation using the four features and the class attribute.

Table 4.1: First five SMS instances for 'hamsAndEftScams.csv' corpus

SenderNu-mLen	SenderSaved-As+362626	content	content-Len	smsClass
12	0	Just go for weekly prices. I have 45 cos started my team in gw 2	64	ham
12	0	FNB:-) D OIVER sent N¤ 1100.00. At FNB ATM, press PROCEED, select eWallet Services. PIN 92782, expires at 01:21. If expired, dial *140*392# for new PIN	152	eftScam
12	0	Good morning. No, she will probably get it at your place. Sorry for the late reply I'm at the hospital, working	113	ham
5	0	FNB Credit Card SMS Statement: As at 20 Sep account ..102000 statement balance is 37474.21 and a minimum payment of 1914.00 is due by 15 Oct. Thank You.	153	ham
12	0	FNB:-) K AMON sent N¤ 2000.00. At FNB ATM, press PROCEED, select eWallet Services. PIN 35996, expires at 23:36. If expired, dial *140*392# for new PIN	152	eftScam

The electronic version of the complete 'hamsAndEftScams.csv' and other corpus that were derived from it and referenced in this thesis are available for research purposes

on request. As outlined in the Procedures section, the 'hamsAndEftScams.csv' SMS corpus was converted to an *'ARFF'* format for compatibility with machine learning classifiers. Figure 4.3 depicts the first five data instances from the 'hamsAndEftScams.arff' along with declarations for the corpus name and features datatypes.

hamsAndEftScams.arff

```
1  @relation hamsAndEftScams
2
3  @attribute senderNumLen numeric
4  @attribute senderSavedAs362626 {1,0}
5  @attribute content string
6  @attribute contentLen numeric
7  @attribute smsClass {ham,eftScam}
8
9  @data
10 12,0,"Just go for weekly prices. I have 45 cos started my team in gw 2",64,ham
11 12,0,"FNB:-) D OIVER sent N¤ 1100.00. At FNB ATM, press PROCEED, select eWallet
   Services. PIN 92782, expires at 01:21. If expired, dial *140*392# for new
   PIN",152,eftScam
12 12,0,"Good morning. No she will probably get it at your place. Sorry for the
   late reply I'm at the hospital, working",113,ham
13 5,0,"FNB Credit Card SMS Statement: As at 20 Sep account ..102000 statement
   balance is  37474.21 and a minimum payment of  1914.00 is due by 15 Oct.Thank
   You.",153,ham
14 12,0,"FNB:-) K AMON sent N¤ 2000.00. At FNB ATM, press PROCEED, select eWallet
   Services. PIN 35996, expires at 23:36. If expired, dial *140*392# for new
   PIN",152,eftScam
```

Figure 4.3: First five data instances for 'hamsAndEftScams.arff' corpus

4.3 Extracted Features

The feature extraction, which used *'StringToWordVector'* unsupervised filter, converted the SMS contents to a set of numeric attributes representing words or terms and their occurrences in each SMS instance. A total of 1223 features were extracted from the contents of the 240 SMSes constituting the 'hamsAndEftScams.arff' corpus. The set of all the extracted features are presented in Appendix C. After the feature extractions the 1223 features in addition to the three original features (i.e. *'senderNumLen', 'senderSavedAs362626'and 'contentLen')* formed a set with a total

of 1226 features, which together with the '*smsClass*' attribute defined SMSes in the resultant 'hamsAndEftScamsFtrs.arff' corpus.

Figure 4.4 depicts the first 5 SMS instances along with declarations for the first 7 features in the 'hamsAndEftScamsFtrs.arff' corpus.

```
@relation 'hamsAndEftScams-weka.filters.unsupervised.attribute

@attribute senderNumLen numeric
@attribute senderSavedAs362626 {1,0}
@attribute contentLen numeric
@attribute ***il numeric
@attribute *101*27$ numeric
@attribute *140*295$ numeric
@attribute *140*392$ numeric
@data
{0 12,1 0,2 64,59 1,110 1,305 1,412 1,433 1,452 1,460 1,485
1,492 1,519 1,612 1,739 1,860 1,893 1,995 1,1226 ham}
{0 12,1 0,2 152,6 1,9 1,11 1,68 1,198 1,199 1,331 1,384
1,388 1,389 1,411 1,412 1,489 1,631 1,651 1,720 1,737 1,746
1,812 1,815 1,819 1,1062 1,1151 1,1170 1,1200 1,1226 eftScam}
{0 12,1 0,2 113,198 1,412 1,429 1,438 1,473 1,485 1,513
1,538 1,566 1,602 1,640 1,723 1,745 1,779 1,823 1,851 1,907
1,1009 1,1023 1,1045 1,1226 ham}
{0 5,1 0,2 153,9 1,26 1,41 1,55 1,60 1,68 1,101 1,144 1,150
1,176 1,192 1,198 1,209 1,243 1,254 1,310 1,351 1,411 1,511
1,590 1,654 1,656 1,710 1,816 1,845 1,861 1,902 1,1044
1,1226 ham}
```

Figure 4.4: SMS instances and features for 'hamsAndEftScamsFtrs.arff'

An elaborate description of the data fields in figure 4.4 is presented in the **Data** Analysis and Discussion chapter.

4.4 Features for Optimal SMS Classification

As described in the Procedures section, selecting optimal classification features first required determining the optimal IG_{tshld}. The optimal IG_{tshld} was determined by employing feature selection using different IG_{tshld} values and subsequently using the

selected features to test the three classifiers performance. Table 4.2 shows the total number of features selected using different IG_{tshld} values.

Table 4.2: IG_{tshld} values versus the total number of selected features

IG Threshold Value	Total No. of Selected Features
0.000	120
0.025	49
0.050	26
0.075	23
0.100	22

Figure 4.5 shows the average percentage CA for the three classifier models during 10-fold CV experiments to determine the optimal IG_{tshld}.

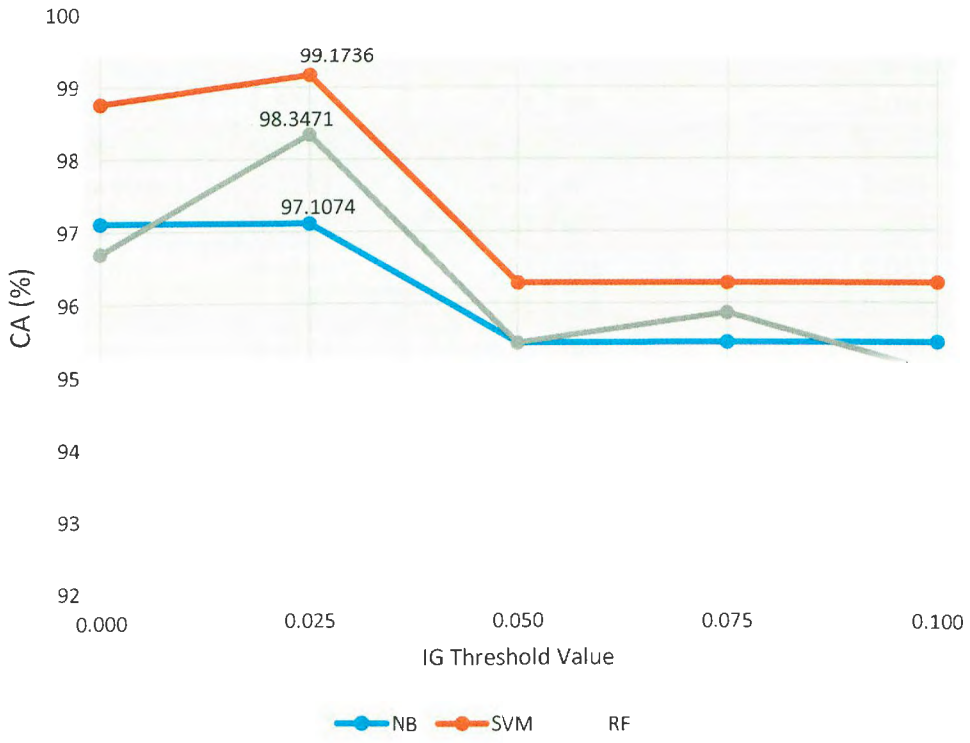


Figure 4.5: IG_{tshld} values versus classifiers' CA

Figure 4.5 shows that the 49 features selected with IG_{tshld}=0.025 allowed each of the classifier models to either maintain or achieve its highest observed CA. The highest achieved CA translates to 97.1%, 99.2% and 98.3% for NB, SVM and RF respectively. The 49 optimal features are listed in Table 4.3 along with their respective index numbers from the 1226 features set and their computed *IG* values.

Table 4.3: Optimal classification features and their *IG* values

Index Number	Feature	Feature IG value	Index Number	Feature	Feature IG value
385	ewallet	0.6416	514	it	0.0519
652	na¤	0.6158	1	senderNumLen	0.0475
332	dial	0.615	1209	querries	0.0437
7	*140*392#	0.6134	2	senderSavedAs362626	0.0437
816	sent	0.6028	1056	06129922	0.0437
3	contentLen	0.5421	177	and	0.0424
10	00	0.5392	927	to	0.0401
813	select	0.5372	51	16hrs	0.0393
412	fnb	0.5309	1016	with	0.0368
389	expired	0.5233	671	on	0.0368
738	press	0.5233	1046	your	0.035
747	proceed	0.5233	657	of	0.035
820	services	0.5233	1010	will	0.035
200	atm	0.498	252	can	0.0331
721	pin	0.498	174	am	0.0331
632	new	0.4754	578	me	0.0313
490	f	0.365	8	*140*999#	0.029
390	expires	0.3431	190	are	0.0277
199	at	0.2779	686	or	0.0277
413	for	0.2682	1091	2350	0.026
1045	you	0.132	1072	1600	0.026
1109	362626	0.1281	1068	14	0.026
908	the	0.0822	726	please	0.0259
486	i	0.0686	493	in	0.0259
972	valid	0.0626			

From Table 4.3, the three features, *'senderNumLen'*, *'senderSavedAs362626'* and *'contentLen'* originally used to represent the raw SMSes are also among the selected

features for optimally classifying ham and EFT scam SMSes. Additionally, the table shows that *'contentLen'* with an *IG*=0.5421 lies in the upper quartile, while *'senderNumLen'* and *'senderSavedAs362626'* with *IG*=0.0475 and *IG*=0.0437 respectively lies in the middle of the selected features ranking.

4.5 Classifier Models Pre-evaluation

The pre-evaluation was done based on the average percentage CA. The pre-evaluation results are shown in the following Figure 4.6.

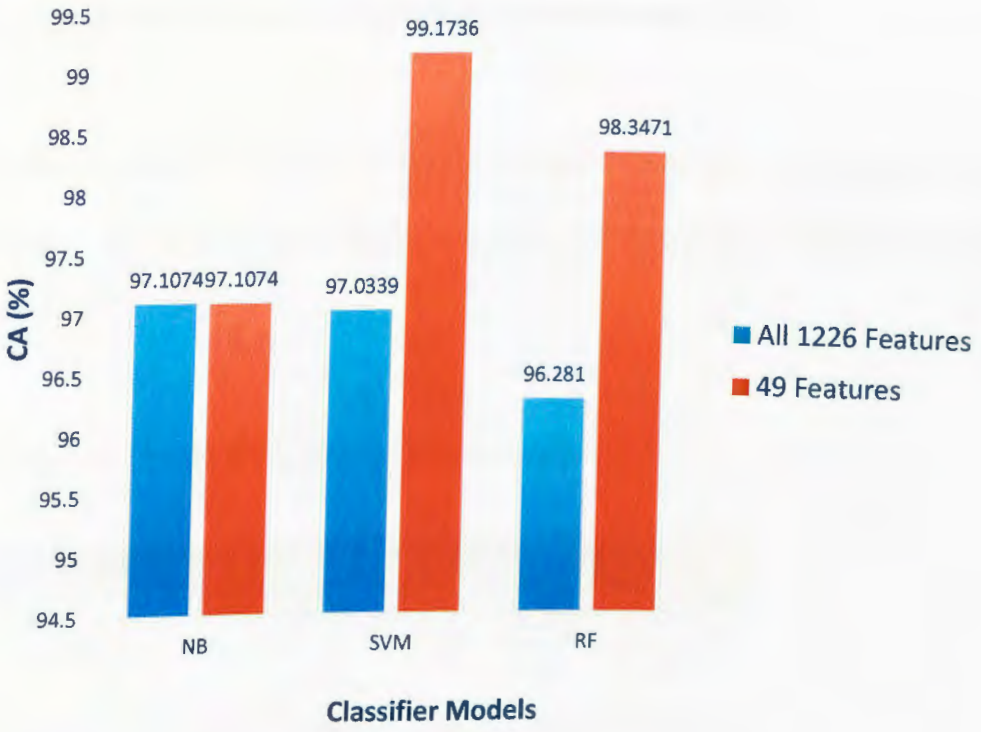


Figure 4.6: Classifier models pre-evaluation results

4.6 Classifier Models Evaluation

The classifier models evaluation was done following a step by step 10-fold cross-validation and used the confusion matrix TP, TN, FP and FN as basic evaluation

metrics. These metrics were defined with respect to ham and spam SMSes in the Classifier Evaluation Metrics section of the Literature Review. Prior to presenting the three classifier evaluation results using these metrics with respect to ham and EFT scam SMSes, it is imperative to redefine them in that context to prevent possible obscurities. The four metrics could be redefined as follows:

TP: the number of EFT scam SMSes that are correctly classified

TN: the number of ham SMSes that are correctly classified

FP: the number of ham SMSes falsely classified as EFT scam SMSes

FN: the number of EFT scam SMSes falsely classified as ham SMSes

Table 4.4a, 4.4b and 4.4c outlines the TP, TN, FP and FN results for each of the 10-fold CV training and testing iteration for the NB, SVM and RF classifier models respectively.

Table 4.4a: NB 10-fold CV confusion matrix results

Training-testing Iteration	TP	TN	FP	FN
1	19	5	0	0
2	18	5	0	1
3	17	6	0	1
4	18	5	1	0
5	17	6	0	1
6	18	6	0	0
7	17	6	0	1
8	18	6	0	0
9	18	6	0	0
10	17	6	0	1

Table 4.4b: SVM 10-fold CV confusion matrix results

Training-testing Iteration	TP	TN	FP	FN
1	19	5	0	0
2	19	5	0	0
3	18	6	0	0
4	18	4	2	0
5	18	6	0	0
6	18	6	0	0
7	18	6	0	0
8	18	6	0	0
9	18	6	0	0
10	18	6	0	0

Table 4.4c: RF 10-fold CV confusion matrix results

Training-testing Iteration	TP	TN	FP	FN
1	19	5	0	0
2	19	5	0	0
3	17	6	0	1
4	18	4	2	0
5	18	6	0	0
6	18	6	0	0
7	18	6	0	0
8	18	6	0	0
9	17	6	0	1
10	18	6	0	0

The average TP, TN, FP and FN values for each of the three classifier's 10-fold CV training and testing iteration are presented in Table 4.5.

Table 4.5: Classifier models 10-Fold CV average TP, TN, FP and FN values

		10-fold CV Average Values			
		TP= $\frac{1}{10}\sum_{i=1}^{10} TP_i$	TN= $\frac{1}{10}\sum_{i=1}^{10} TN_i$	FP= $\frac{1}{10}\sum_{i=1}^{10} FP_i$	FN= $\frac{1}{10}\sum_{i=1}^{10} FN_i$
Classifier Models	**NB**	17.70	5.70	0.10	0.50
	SVM	18.20	5.60	0.20	0.00
	RF	18.00	5.60	0.20	0.20

Sound analysis and comparison of the classifier models performance require using other evaluation metrics such as FPR, FNR, CA, Precision, Recall and F1-measure in addition to the four basic metrics constituting the confusion matrix. Once more, prior to presenting the computed results in terms of the aforementioned evaluation metrics, these metrics can be redefined as follows in the context of ham and EFT scam SMSes as follows:

FPR: The rate of ham SMSes misclassification as EFT scam SMSes.

FNR: The rate of EFT scam SMSes misclassification as ham SMSes.

CA: The ratio of the number of correctly classified SMSes to the total number of input SMSes to the classifier model.

Precision: The ratio of messages that are classified as EFT scams and are actually EFT scams.

Recall: The ratio of messages that are actually EFT scams and are classified accurately as EFT scams.

F-measure: The harmonic mean of Precision and Recall.

Table 4.6 presents computed values for the three classifier models performances in terms of the just presented evaluation metrics.

Table 4.6: Computed values for classifier models evaluation metrics

		Classifier Models		
		NB	**SVM**	**RF**
Evaluation Metrics	$FPR = \frac{FP}{FP+TN}$	0.017	0.034	0.034
	$FNR = \frac{FN}{FN+TP}$	0.027	0.000	0.011
	$CA = \frac{TP+TN}{TP+FP+TN+FN}$	0.975	0.992	0.983
	$Precision = \frac{TP}{TP+FP}$	0.994	0.989	0.989
	$Recall = \frac{TP}{TP+FN}$	0.973	1.000	0.989
	$F1\ measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$	0.983	0.995	0.989

More descriptive visualisation and analysis of the classifier evaluation results displayed by Table 4.4, 4.5 and 4.6 are presented in the next chapter.

4.7 Summary

A corpus of 240 SMSes was collected from the FNB Namibia m-banking users. This corpus comprised of 184 ham and 56 EFT scam SMSes, respectively. The employed feature extraction process extracted 1223 features from the '*content*' texts of SMSes comprising the corpus, which in addition to '*senderNumLen*', '*senderSavedAs362626*' and '*contentLen*' features that defined raw SMSes formed a 1226 features set. A total of 49 optimal classification features were selected using IG feature selection, that used an IG_{tshld} = 0.025. The 10-fold CV classifiers training and evaluation used SMSes defined using the 49 optimal features. The TP, TN, FP, FN, FPR, FNR, CA, Precision, Recall and F1-measure metrics were used to evaluate the NB, SVM and RF classifier models performances.

The next chapter provide an analysis and discussions of the findings and results presented in this chapter.

5. DATA ANALYSIS AND DISCUSSION

This chapter presents an analysis of the research data and results. It further presents discussions and interpretations of the research findings.

5.1 Collected SMSes, SMS Representation and Feature Extraction

From the four features used to represent raw SMSes, the *'content'* which comprised of the SMS bodies contains most of the SMSes textual information. Such textual information can as well be used in classification of the SMSes. It is, however, difficult to identify which of these textual features are appropriate to classify the SMSes. Feature extraction and feature selection were hence correspondingly required to capture the word and term features from the SMS contents and to determine which ones are appropriate to classify the SMSes efficiently.

A total of 1223 features were extracted from SMS contents and used together with *'senderNumLen'*, *'senderSavedAs+362626'*, *'contentLen'*, and the *'SMSclass'* attribute to define SMSes in the 'hamsAndEftScamsFtrs.arff' corpus. Figure 5.1 provides descriptions explaining data structuring for the 'hamsAndEftScamsFtrs.arff' corpus.

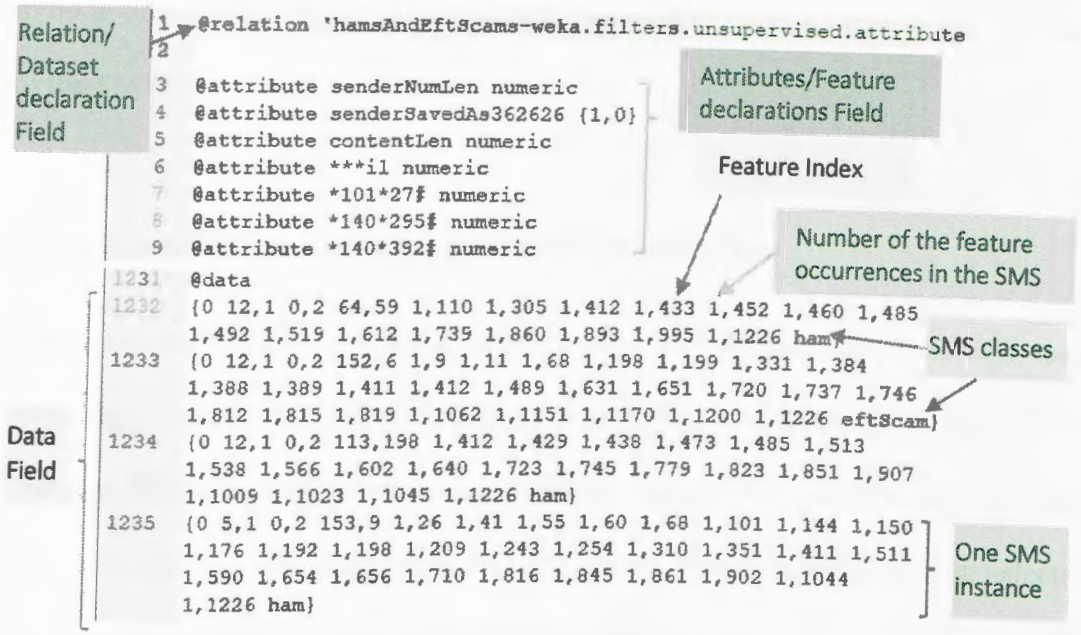


Figure 5.1: Description of SMS representations after feature extraction

From Figure 5.1, the SMS features are represented using indices from 0-1225 assigned in the same order as the features declarations. The 1226th index represents the '*SMSclass*' attribute at the end of each SMS instance. As shown, all SMS instances start and end with an opening and a closing brace, respectively. All the SMSes have '*senderNumLen*', '*senderSavedAs362626*' and '*contentLen*' features, *which* are represented by 0, 1 and 2 indices respectively, as the first three features. These three feature indices are paired with numbers specifying the digits comprising the SMS sender number, whether the sender number is saved as +362626 on the victims' mobile device and the number of characters in the SMS body respectively.

For every SMS instance, the first three features are followed by indices for the extracted features present in that specific SMS paired with the number of its occurrences. Although the indices of the extracted features present in each SMS appear

to be in order from the lowest to the highest, the order plays no role in classification of the SMSes.

5.2 Selection of Features for Optimal SMS Classification

During the process of determining which features allow optimal ham and EFT scam SMS classification, the 1226 features used to define and represent SMSes in the "hamsAndEftScams.arff" corpus were evaluated using *IG* feature selection algorithm. The evaluation computed and ranked the *IG* values for all the features, then selected the features whose *IG* values met the specified IG_{tshld}.

It was showed in Table 4.2 that 120 features were selected using $IG_{tshld} = 0$. This implied that only 120 out of 1226 features have *IG* values greater than zero. The remaining 1106 features have $IG = 0$. This means that by the *IG* feature selection evaluation they offer no significant information towards making a distinction between the ham and EFT scam SMSes. This could be verified by comparing the classifier models percentage CA when all features were used to that when only features selected with $IG_{tshld} = 0$ were used. The comparison is summarised in Figure 5.2.

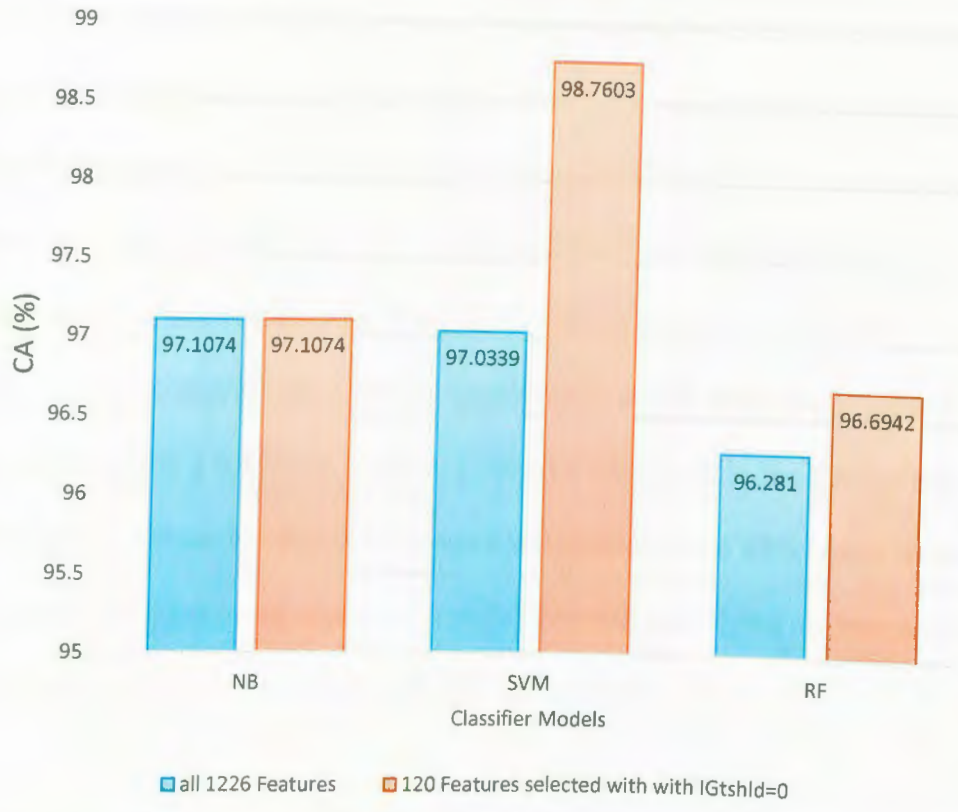


Figure 5.2: CA with 1226 features versus with 120 features

Comparisons in Figure 5.2 shows that the percentage CA for the NB classifier remained at 97.1074%, both when the SMSes representation used all 1226 features and when it only used 120 features selected with $IG_{tshld}=0$. This indicates that the differences in features defining the SMSes did not result in any net change in CA for the NB classifier. The percentage CA for SVM and RF stood at 97.0339% and 96.281%, respectively, when the SMSes representation used all 1226 features but went up to 98.7603% and 96.6942%, respectively, when the SMSes were represented using 120 features selected with $IG_{tshld}=0$. These results are in line with the deductions cited in the Literature Review, that not all text features are significant towards classification, and that the features that are not important to classification often only degrades the classifiers performance.

The employed *IG* feature selection used IG_{tshld} values from the {0.0, 0.025, 0.05, 0.075, 0.10} set. From this set, the $IG_{tshld} = 0.025$ which led to a selection of 49 features emerged as the optimal IG_{tshld}. This was informed by observation that using SMS datasets defined using the 49 features allowed the NB, SVM and RF classifier models to either maintain or reach the highest average CA they have achieved. The NB classifier model maintained its highest 97.1074% CA while SVM and RF models reached 99.1736% and 98.3471% CA, respectively, which were their highest. For SVM, this signified a 0.4133% increase from 98.7603%, while for RF it showed a 1.6529% increase from 96.6942% CA recorded with the use of all features which by the *IG* evaluation contributed useful information towards classifying the ham and EFT scam SMSes (i.e. 120 features selected with $IG_{tshld} = 0.0$).

5.3 Classifier Models Pre-evaluation

As acknowledged in the Research Methods and the Research Findings chapters, the pre-evaluation experiments attempted to validate the necessity to use optimal features to classify ham and EFT scam SMSes as opposed to using all features. As demonstrated by Figure 4.6 in the Research Findings, with an exception to the NB classifier which maintained its highest recorded CA, both SVM and RF classifier models showed better or improved CA when optimal features were used instead of all 1226 features. In this respect, the SVM classifier model CA showed a 2.1397% improvement from 97.0339% to 99.1736%, while RF showed a 2.0661% increase from 96.281% to 98.3471%. These results essentially provided the required validation.

5.4 Classifier Models Evaluation

The three classifier models evaluation, which followed a step by step 10-fold CV approach, did not only allow comparisons of the models average performances but also permitted close examination and analysis of their behaviours during each 10-fold CV training and testing iteration. Sections 5.4.1 and 5.4.2 present an analysis and discussions of the evaluation results.

5.4.1 Confusion Matrix Evaluation Metrics

The three classifier model results for the 10-fold CV training and testing iterations in terms of confusion matrix TP, TN, FP, and FN values are presented in point 5.4.1.1 through 5.4.1.3.

5.4.1.1 Naïve Bayes

Table 4.4a presented the TP, TN, FP and FN values for each of the 10-fold CV training and testing iterations for the NB classifier model. The presented results show that four of the iterations (i.e. 1, 6, 8 and 9) did not produce any misclassification. The FP and FN are zeroes for the aforementioned iterations. The results further indicate that the five iterations (i.e. 2, 3, 5, 7 and 10) all have FN=1, meaning that each of them has resulted in one wrongly classified EFT scam SMS. From the table, only the 4th iteration had misclassified a ham SMS, having a FP =1.

While the results clearly show that the NB classifier have very low FP values (i.e. only one iteration have a non-zero FP), they also show that most of the iterations (i.e. 5) have an EFT scam SMS misclassified as a ham SMS (i.e. FN=1). This infers that the NB classier model correctly predicted ham SMSes class more effectively compared to

EFT scam SMSes class. The average TP, TN, FP and FN values for the 10-fold CV training and testing iterations stands at 17.7, 5.7, 0.1 and 0.5 respectively, in conformity with the preceding explanations.

5.4.1.2 Support Vector Machine

Table 4.4b presented the TP, TN, FP and FN values for the 10-fold CV training-testing iterations for the SVM classifier model. The table shows that from the 10 iterations, only one resulted in an SMS misclassification. This was the 4^{th} iteration which has FP = 2. The remaining 9 iterations all have FP = 0 and FN = 0. The corresponding average values for TP, TN, FP, and FN metrics are 18.2, 5.6, 0.2 and 0.0 respectively. An essential deduction could be made from the average FN= 0.0, whose value implies that the SVM classifier model correctly classified all EFT scam SMSes that comprised the testing folds.

5.4.1.3 Random Forest

Table 4.4c outlined the 10-fold CV training and testing iterations TP, TN, FP and FN results for the RF classifier model. The results show that only the 3^{rd}, 4^{th} and 9^{th} iterations have produced SMS misclassifications. Both the 3^{rd} and 9^{th} iterations have FN = 1, meaning that each has a misclassified EFT scam SMS. The 4^{th} iteration on the other hand have FP = 2, having misclassified 2 ham SMSes. The average TP, TN, FP and FN for the RF classifier model stands at 18.0, 5.6, 0.2 and 0.2 respectively. Examining these average values gives a glimpse that the RF model performance lies somewhere in between that of NB and SVM.

In relation to each other, the average TP and TN values for the three models shows that with respect to correctly classifying ham SMSes, the NB model perform the best (i.e. TN =5.70), while SVM and RF models performs on par (i.e. both with TN=5.60). In terms of the capability to correctly classify EFT scam SMSes, the SVM model performed the best with TP = 18.20, followed by RF with TP=18.0, then lastly the NB with TP = 17.70. For this study's context, the models ability to correctly classify EFT scam SMSes carries more weight compared to its ability to correctly classify ham SMSes. This is the case because the key purpose for employing classification is to detect the EFT scam SMSes in order to address the associated problem.

The three models performance in terms of the average FP and FN values could be better expressed and interpreted in terms of FPR and FNR, respectively. These two rates are discussed in the next section.

5.4.2 Other Evaluation Metrics

Although the TP, TN, FP and FN values for the three classifier models discussed in section 5.4.1 give a preview of how the models performed, these evaluation metrics are merely adequate to make a comprehensive analysis and comparisons of the performances. Analysing the three models performances using the remainder of the evaluation metrics such as FPR, FNR, CA, Precision, Recall and F1-measure can allow a more elaborate discussion and interpretation of their strengths and weaknesses. Section 5.4.2.1 looked at the three classifier models' performance in terms of FPR and FNR evaluation metrics, while section 5.4.2.2 looked at it from the perspective of the CA, Precision, Recall and F1-measure evaluation metrics.

5.4.2.1 FPR and FNR

Generally, it is desirable to have the FPR and FNR for a classifier model as low as possible. If a classifier model has lower FPR and FNR, it means that it misclassifies data instances at a lower rate. Figure 5.3 compares the FPR and FNR for the three classifier models.

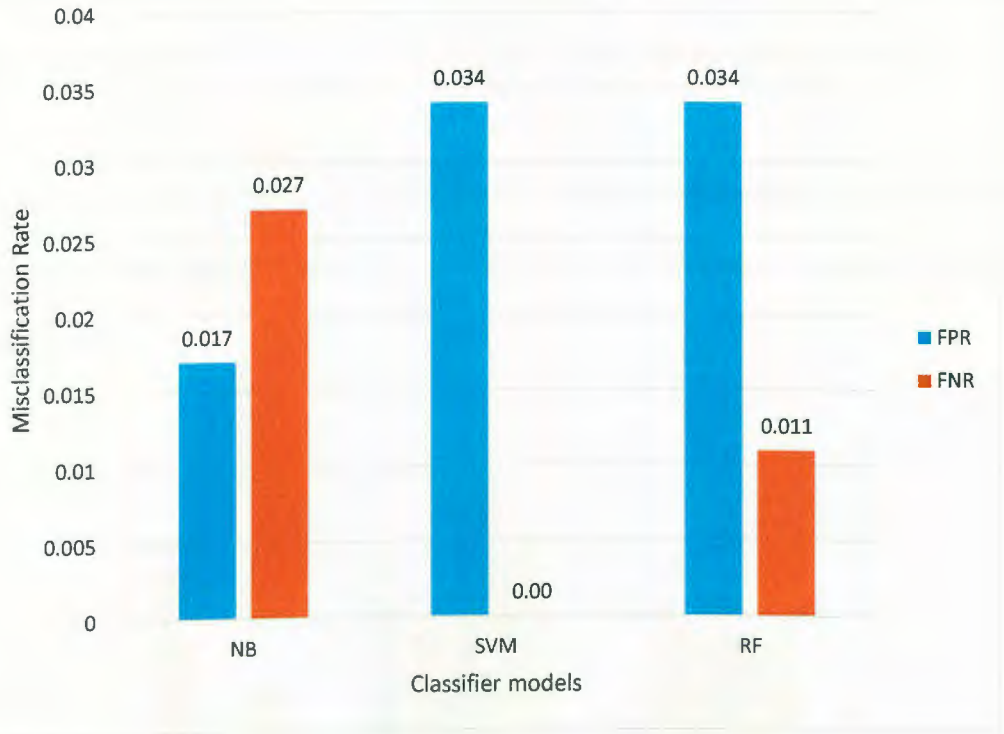


Figure 5.3: Classifier models FPR and FNR

The FPR for the NB, SVM and RF classifier models stands at 0.017, 0.034 and 0.034, respectively. The lowest FPR for the NB suggests that it is the most efficient among the three models with respect to correctly classifying ham SMSes. Both SVM and RF models have the same FPR =0.034, meaning that they misclassify ham SMSes at the same rate.

The FNR stands at 0.027 for NB, 0.0 for SVM and 0.011 for RF classifier model. These values imply that out of the three models, SVM performs the best class prediction for EFT scam SMSes. Having a FNR=0.0 means that the SVM have not misclassified a single EFT scam SMS over the 10-fold CV training and testing iterations. The RF model has FNR=0.011, which is the second lowest amongst the three models. This put the RF model second-best with respect to the rate of correctly classifying the EFT scam SMSes. The NB has the highest FNR = 0.027, demonstrating that it misclassified EFT scam SMSes at a higher rate than the other two models.

Figure 5.4 gives a graphical presentation of the models' performances in terms of CA, Precision, Recall and F1-measure evaluation metrics to facilitate further analysis and comparisons.

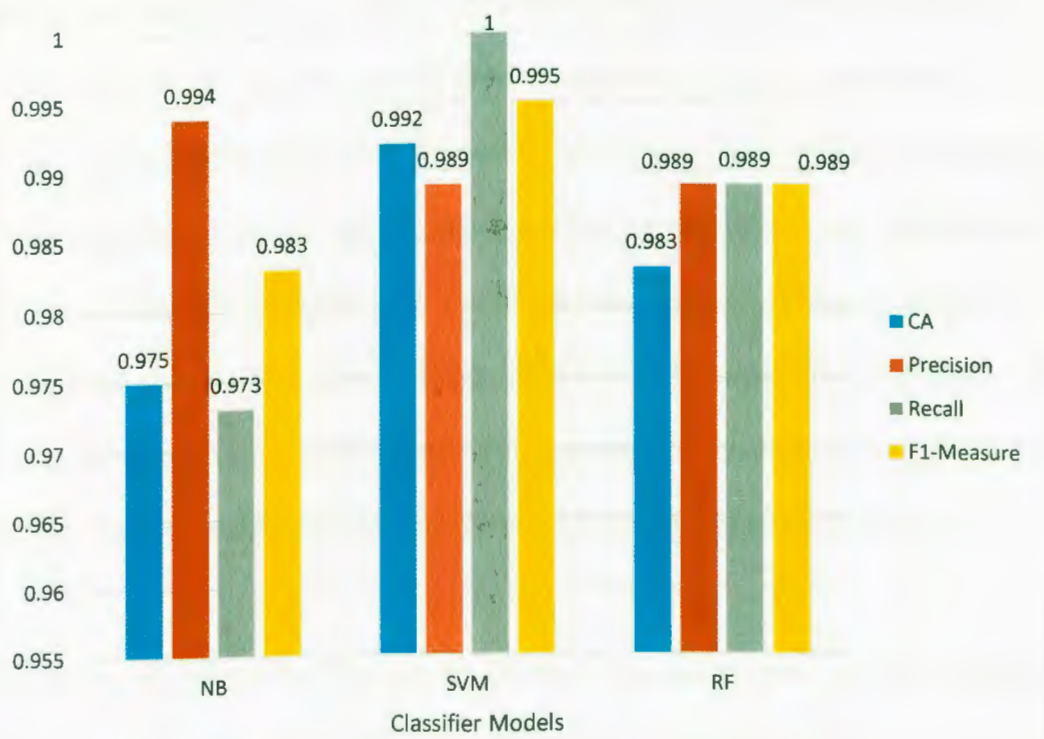


Figure 5.4: Classifier models CA, Precision, Recall and F1-measure

The NB, SVM and RF classifier models have a 0.975, 0.992 and 0.983 CA, respectively. These values show that on overall, the SVM model makes more correct predictions of SMS classes, followed by RF, then the NB model which comes last.

The NB model showed the highest Precision = 0.994 compared to the SVM and RF which both have Precision = 0.989. The Recall values for the NB, SVM and RF are 0.973, 1.0 and 0.989, respectively. SVM has the highest possible Recall value, meaning that it had correctly classified all EFT scam SMSes, agreeing with the interpretations presented in section 5.4.1. The Recall results also shows that the RF model has the second-best performance with respect to correctly classifying EFT scam SMSes while the NB model comes last.

From the discussions in the two previous paragraphs, the three classifier models have contrasting performance with respect to Precision and Recall (i.e. NB have the highest Precision than that for the SVM and RF models which are equal, but the lowest Recall compared to that for the SVM and RF models, which are highest and second-highest respectively). This can easily bring confusion when trying to determine which of the three classifier models gives the best overall performance with respect to classifying both the ham and the EFT scam SMSes. This brings the necessity to analyse the Precision and Recall values with respect to their respective equations (i.e. equation 2.8 and 2.9) in order to explain the reasons for the contrasting model performances.

From the two equations, the differences between the two metrics which is in their denominators is represented by FP in Precision and FN in Recall. As discussed in the previous section the NB model have a superior performance with respect to correctly

classifying the ham SMSes (i.e. low FP) while the SVM and RF outperform it with respect to correctly classifying the EFT as cam SMSes (i.e. low FN). This explains the observed contrasting performances of the three models in terms of Precision and Recall values.

The F1-measure evaluation metric, which represents the harmonic mean of Precision and Recall, helps ease comparisons of different classifier models in scenarios such as the one just described. The F1-measure values for the NB, SVM and RF models are 0.983, 0.995 and 0.989, respectively. This sum up the overall performances for the three classifier models, revealing that the SVM model performed the best, followed by RF, then lastly the NB.

5.5 Summary

Evaluating the three classifiers indicated that the NB model performed the best with respect to correctly classifying ham SMSes, followed by SVM and RF with similar performances. From the perspective of the classifiers capability to detect EFT scam SMSes, the SVM model performed the best, followed by RF, then lastly NB. The overall evaluation result implies that the SVM classifier model is the most suited for ham and EFT scam SMS classification application, followed by RF, then lastly NB.

The next chapter looks at recommendations and concludes the study.

6. RECOMMENDATIONS AND CONCLUSION

This chapter presents recommendations drawn based on data analysis, discussions and interpretations of the research findings. It expounds whether the research objectives were met and shed lights on possible future works, before concluding with a brief summary of the entire research work.

6.1 Recommendations

The study progression from the introduction and review of similar works, through to data analysis and discussions saw various key observations and inferences made. Such observations and inferences led to the formulation of various recommendations. The subsequent Section 6.1.1 through 6.1.4 summarises the main recommendations drawn in doing this research.

6.1.1 SMS Collection

Due to the inexistence of appropriate public corpora for ham and EFT scam SMSes, this study collected its own SMS datasets from public Facebook groups and from individual volunteers. Although some of the shortcomings associated with collecting SMS datasets using such approaches were expected as per revelations from the reviewed literature, the full extent of such shortcomings was seemingly underestimated. The collection of particularly the EFT scam SMSes from individuals turned out to be very slow then initially anticipated, pushing the data collection duration from the planned 4 weeks to nearly 20 weeks, leading to the extension of the study duration. From this observation, this study recommends that prospective scholars only consider the method of collecting EFT scam SMSes from volunteers

viable when they can afford to dedicate enough time to the data collection. If they do not have sufficient time to collect the EFT scams or related SMSes, then they should consider exploring avenues such as providing incentives to encourage volunteer to contribute to the datasets. The Literature Review chapter had revealed that such methods have been used in similar works to fast track the process of collecting spam SMSes. This study, however, opted to rather extend the SMS collection time, due to the fear that providing incentives might encourage individuals to provide false EFT scam SMSes.

While collecting the EFT scam SMSes from volunteers was very slow, extracting them from Facebook user posts was fairly fast. With the continuing explosive growth in popularity, similar works should not overlook the Social media as a possible data collection tool or source. Having observed the lack of public corpora suitable to study machine learning classification of ham and EFT scam SMSes, the datasets collected in this study would be available on request for research purposes.

6.1.2 Feature Extraction

This study utilised '*StringToWordVector*' unsupervised filter to extract words or terms features from contents of SMSes that comprised the used 'hamsAndEftScams.arff' corpus. Although the research experiments showed that this method has extracted features that allowed building classifier models that performed fairly well, there are other feature extraction methods that could as well be used. Other works can consider using different feature extraction methods and compare the resultant classifier models' performance to those in this study.

6.1.3 Feature Set and Optimal Classification Features

A total of 1223 features were extracted from the 240 SMS contents and used together with 'senderNumLen', 'senderSavedAs362626', 'contentLen' and 'smsClass' to represent the SMSes. An application that safeguards m-banking users from EFT scams by employing machine learning classification had to run on users' mobile device. Hence, it will not be ideal for such an application to store and execute SMS classification using such a large number of features considering the limited storage and processing resource for mobile phones.

The *IG* feature selection process identified a total of 49 features, which, when used to define the SMSes, allowed the classifier models to predict classes with a better CA compared to when all features were used. Deducing from the reduced features set and the associated improved classifiers performance, this study recommends that the works with classifiers models intended for mobile devices always consider employing features selection. Furthermore, apart from *IG* feature selection, other methods like correlation can also be used for the same application. Similar works can also explore the use of such methods.

6.1.4 Classifiers' Suitability to Detect EFT Scam SMSes

The overall evaluation results of the three classifiers show that the SVM classifier model performs more accurate class predictions for the ham and EFT scam SMSes compared to the RF and NB models. The particular strengths for the SVM model comes with respect to its demonstrated capability to correctly classify or detect EFT scam SMSes. Having produced an average FN=0 meant that it has not misclassified a single EFT scam SMS. Hence, if one is looking to implement an application that uses

machine learning classification in a context that requires giving more priority to scam detections such as this study, then it is imperative that they first look into using the SVM classifier. RF classifier can be considered second before NB as a last option.

6.2 Future Works

The envisaged future works extension to this study will involve investigating the possibility to tackle the research problem from a multiclass classification perspective. The possibility of extending the used evaluation metrics to include others such as Area Under the ROC Curve and the models' speed will also be examined. Furthermore, the methods, procedures, observations and recommendations drawn in this study would be used to guide a development of an actual mobile application that employ SVM machine learning classifier to detect EFT scam SMSes on user devices.

6.3 Conclusion

The study collected a corpus of 184 ham and 56 EFT scam SMSes and used WEKA data mining platform to perform the experiments. The raw SMSes were defined using four features and a class attribute. This followed inferences drawn from the surveyed literature and an analysis of ham and EFT scam SMS characteristics.

Filtering was employed to extract words and term features from the SMS *contents* so that they can be used for classification together with the other features used to define the raw SMSes. After the extraction, all the features were evaluated using IG algorithm, leading to the determining of 49 optimal features that were used to define SMS instances prior to the classifiers training and testing or evaluation.

The training phases involved learning the NB, SVM and RF classifiers to predict the ham and EFT scam SMS classes. This saw the fulfilment of the Research Objective a). During testing, the trained classifier models' ability to predict classes for SMSes comprising the test folds was evaluated based on the TP, TN, FP, FN, FPR, FNR, CA, Precision, Recall and F1-measure metrics. This saw the fulfilment of the final Research objective b).

Evaluating the three models showed that the NB can correctly classify ham SMSes more effectively than the SVM and RF which performed on par in that regard. The SVM showed a superior performance with respect to the capability to correctly classify or detect EFT scam SMSes, followed by RF then lastly NB. The overall evaluation results showed that the SVM classifier model is the most efficient with respect to classifying both ham and EFT scam SMSes, followed by RF then lastly NB. This is affirmed by the F1-measure values of the three models, which summarises their performances, taking into account both their Precision and Recall values.

The envisaged future work as an extension to this study will consider developing a mobile application that implements an SVM classifier model to allow m-banking user devices detect EFT scam SMSes.

REFERENCES

Abdulhamid, S. M., Abd Latiff, M. S., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., & Herawan, T. (2017). A Review on Mobile SMS Spam Filtering Techniques. *IEEE Access*, *5*, 15650–15666.

Ahmed, I., Guan, D., & Chung, T. C. (2014). SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset. *International Journal of Machine Learning and Computing*, *4*(2), 183–187.

Akbari, F., & Sajedi, H. (2015). SMS spam detection using selected text features and Boosting Classifiers. *2015 7th Conference on Information and Knowledge Technology (IKT)*, 1–5.

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. *Journal of Applied Mathematics*, *2014*, 1–6.

Almeida, T. A., Gómez, J. M., & Yamakami, A. (2011). Contributions to the Study of SMS Spam Filtering: New Collection and Results. *2011 ACM Symposium on Document Engineering*, 259--262. Mountain View.

Aragão, M. V. C., Frigieri, E. P., Ynoguti, C. A., & Paiva, A. P. (2016). Factorial design analysis applied to the performance of SMS anti-spam filtering systems. *Expert Systems with Applications*, *64*, 589–604.

Arde, A. (2012). EFT scammers fake their proof of payment to dupe you, the seller. *IOL Personal Finance*. Retrieved September 26, 2019, from Independent Online (IOL South Africa) website: https://www.iol.co.za/personal-finance/eft-scammers-fake-their-proof-of-payment-to-dupe-you-the-seller-1209136

Azhagusundari, B., & Thanamani, A. S. (2013). Feature Selection based on Information Gain (University of Houston; Vol. 2).

Bouckaert, R. R., Frank, E., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2002). *WEKA Manual for Version 3-7-2* (Vol. 11).

Chen, T., & Kan, M.-Y. (2012). Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, *47*(2), 299–335.

Choudhary, N., & Jain, A. K. (2017). Towards filtering of SMS spam messages using machine learning based technique. *International Conference on Advanced Informatics for Computing Research*, *712*, 18–30.

Christopher, N., & Kar, S. (2018). Mobile wallet scam: As next-gen spenders go cashless, e-wallet scamsters too are getting creative. Retrieved October 5, 2018, from Economic Times website: https://tech.economictimes.indiatimes.com/news/internet/as-next-gen-spenders-go-cashless-e-wallet-scamsters-too-are-getting-creative/63889870

Cormack, G. V. (2008). Email Spam Filtering: A Systematic Review. *Foundations and Trends® in Information Retrieval*, *1*(4), 335–455.

Donges, N. (2019). The Random Forest Algorithm: A Complete Guide. Retrieved May 24, 2020, from Builtin website: https://builtin.com/data-science/random-forest-algorithm

Erongo. (2016). Crooks out to empty wallets. Retrieved October 6, 2018, from Erongo News website: http://www.erongo.com.na/news/crooks-out-to-empty-wallets/?

Ezpeleta, E., Garitano, I., Zurutuza, U., & Hidalgo, J. M. G. (2017). Short Messages Spam Filtering Combining Personality Recognition and Sentiment Analysis.

International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, *25*(Suppl. 2), 175–189.

George, O. (2019). Common Classification Model Evaluation metrics. – Towards Data Science. Retrieved August 20, 2019, from Towards Data Science website: https://towardsdatascience.com/common-classification-model-evaluation-metrics-2ba0a7a7436e

Gov Page SA. (2017). General Scams Warnings. Retrieved September 26, 2019, from Gov Page website: https://www.govpage.co.za/general-scams-warnings/fake-payment-sms-email-confirmation

Günal, S. (2012). Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering and Computer Sciences*, *20*(SUPPL.2), 1296–1311.

Hedieh, S., Parast, G. Z., & Akbari, F. (2016). SMS Spam Filtering Using Machine Learning Techniques: A Survey. *Machine Learning Research*, *1*(1), 1–14.

Hidalgo, J. M. G., Bringas, G. C., Sánz, E. P., & García, F. C. (2006). Content based SMS spam filtering. *Proceedings of the 2006 ACM Symposium on Document Engineering - DocEng '06*, 107.

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11.

Jane, D. (2019). Crime in the City Namibia. Retrieved December 19, 2019, from Facebook website: https://www.facebook.com/groups/616819348371856/permalink/2462372400483199/

John, D. (2017). Ewallet scams. Retrieved from Facebook website: https://www.facebook.com/search/top/?q=ewallet scams&epa

Junaid, M. B., & Farooq, M. (2011). Using evolutionary learning classifiers to do MobileSpam (SMS) filtering. *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation - GECCO '11*, 1795.

Kaya, Y., & Ertuğrul, Ö. F. (2016). A novel feature extraction approach in SMS spam filtering for mobile communication: one-dimensional ternary patterns. *Security and Communication Networks*, *9*(17), 4680–4690.

Khorsi, A. (2007). An Overview of Content-Based Spam Filtering Techniques Bayesian Classifier. *Informatica*, *31*, 269–277.

Mahmoud, T., & Mahfouz, A. (2012). SMS Spam Filtering Technique Based on Artificial Immune System. *International Journal of Computer Science Issues*, *9*(2), 589–597.

Mishra, A. (2018). Metrics to Evaluate your Machine Learning Algorithm. Retrieved August 20, 2019, from Towards Data Science website: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Mujtaba, G., & Yasin, M. (2014). SMS Spam Detection Using Simple Message Content Features. *Journal of Basic Applied Scientific Research*, *4*(4), 275–279.

Nagel, E. (2015). Watch out for these common payment scams – Gumtree Blog. Retrieved September 26, 2019, from Gumtree website: https://blog.gumtree.co.za/watch-out-for-these-common-payment-scams/

Nuruzzaman, M. T., Lee, C., & Choi, D. (2011). Independent and personal SMS spam

filtering. *Proceedings - 11th IEEE International Conference on Computer and Information Technology, CIT 2011*.

Reaves, B., Blue, L., Tian, D., Traynor, P., & Butler, K. R. B. (2016). Detecting SMS Spam in the Age of Legitimate Bulk Messaging. *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks - WiSec '16*, 165–170.

Rohith, G. (2018). Support Vector Machine — Introduction to Machine Learning Algorithms. Retrieved August 19, 2019, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Shahi, T. B., Yadav, A., Shahi, T. B., & Yadav, A. (2014). Mobile SMS Spam Filtering for Nepali Text Using Naive Bayesian and Support Vector Machine. *International Journal of Intelligence Science*, *04*(01), 24–28.

Shaikh, A. A., & Karjaluoto, H. (2015). Mobile banking adoption: A literature review. *Telematics and Informatics*, *32*(1), 129–142.

Shirani-Mehr, H. (2012). SMS Spam Detection using Machine Learning Approach. *Tech. Rep., Stanford University*, 1–4.

Suleiman, D., & Al-Naymat, G. (2017). SMS Spam Detection using H2O Framework. *Procedia Computer Science*, *113*, 154–161.

Tan, H., Goharian, N., & Sherr, M. (2012). Identifying the Pertinent Features of SMS Spam. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '12*, 1175.

Tekerek, A. (2018). Support vector machine based spam SMS detection. *Journal of*

Polytechnic, *22*(3), 779–784.

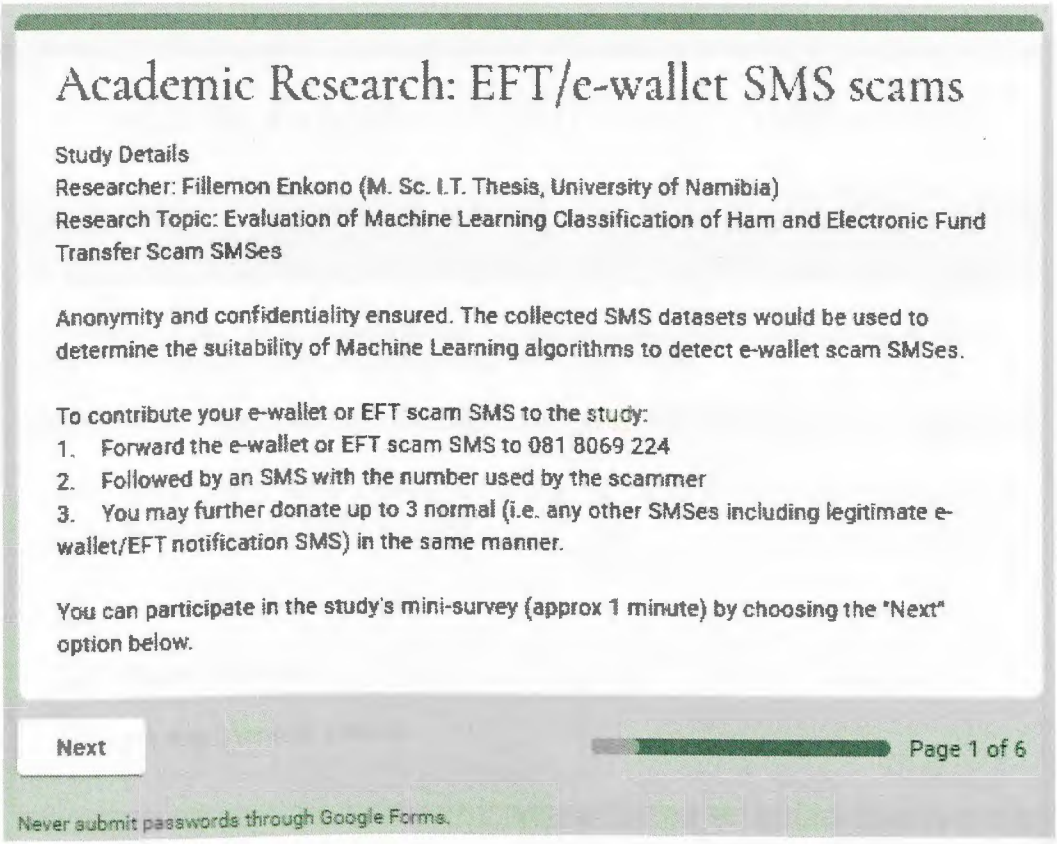
Uysal, A. K., Gunal, S., Ergin, S., & Sora Gunal, E. (2013). The Impact of Feature Extraction and Selection on SMS Spam Filtering. *Electronics and Electrical Engineering*, *19*(5).

Wagiet, R. (2012). SMS banking scam exposed. Retrieved September 26, 2019, from Eyewitness News website: https://ewn.co.za/2012/08/15/Latest-SMS-banking-scam

Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., & Naik, V. (2011). SMSAssassin. *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications - HotMobile '11*, 1.

Yan, Z., Zhang, W., Kantola, R., & Chen, L. (2015). TruSMS: A trustworthy SMS spam control system based on trust management. *Future Generation Computer Systems*, *49*, 77–93.

APPENDIX A - STUDY DETAILS GOOGLE FORM



Academic Research: EFT/e-wallet SMS scams

Study Details
Researcher: Fillemon Enkono (M. Sc. I.T. Thesis, University of Namibia)
Research Topic: Evaluation of Machine Learning Classification of Ham and Electronic Fund Transfer Scam SMSes

Anonymity and confidentiality ensured. The collected SMS datasets would be used to determine the suitability of Machine Learning algorithms to detect e-wallet scam SMSes.

To contribute your e-wallet or EFT scam SMS to the study:
1. Forward the e-wallet or EFT scam SMS to 081 8069 224
2. Followed by an SMS with the number used by the scammer
3. You may further donate up to 3 normal (i.e. any other SMSes including legitimate e-wallet/EFT notification SMS) in the same manner.

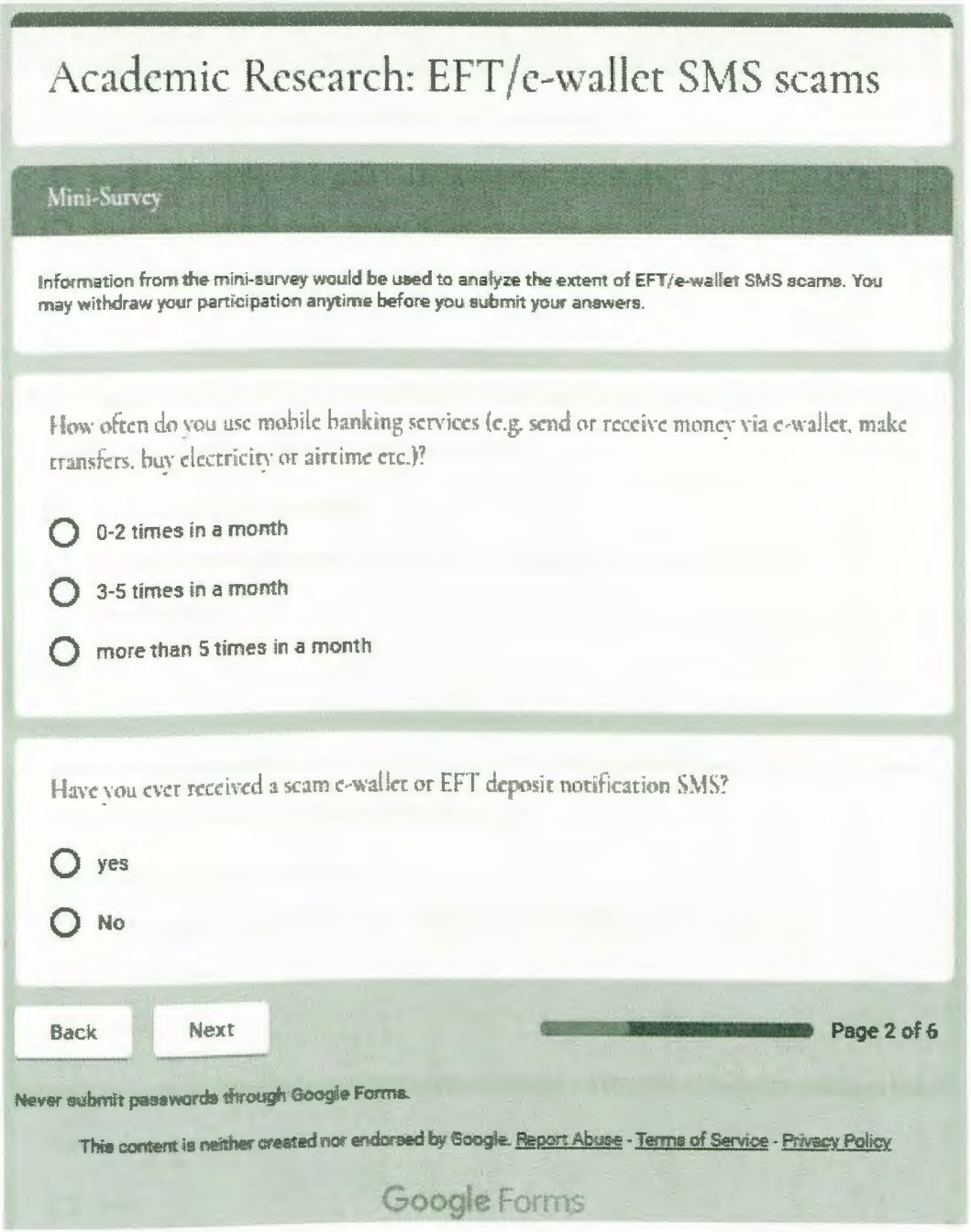
You can participate in the study's mini-survey (approx 1 minute) by choosing the "Next" option below.

Next

Page 1 of 6

Never submit passwords through Google Forms.

APPENDIX B - RESEARCH MINI-SURVEY QUESTIONS



Academic Research: EFT/e-wallet SMS scams

Mini-Survey

Information from the mini-survey would be used to analyze the extent of EFT/e-wallet SMS scams. You may withdraw your participation anytime before you submit your answers.

How often do you use mobile banking services (e.g. send or receive money via e-wallet, make transfers, buy electricity or airtime etc.)?

- 0-2 times in a month
- 3-5 times in a month
- more than 5 times in a month

Have you ever received a scam e-wallet or EFT deposit notification SMS?

- yes
- No

Back Next Page 2 of 6

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Google Forms

Academic Research: EFT/e-wallet SMS scams

If 'yes' to the previous answer:

How did you think the scammers obtained your mobile number?

- ○ From my goods/services sales posts I put up on social media
- ○ From my comments on advertised items on social media
- ○ I do not know
- ○ Other: _____

What do you think the scammers were trying to do?

- ○ Trick me to send them money
- ○ Pretend to have paid for the service/product I was selling, so that they steal it
- ○ I do not Know
- ○ Other: _____

What do you think the scammers were trying to do?

- ○ Trick me to send them money
- ○ Pretend to have paid for the service/product I was selling, so that they steal it
- ○ I do not Know
- ○ Other: _____

Did the scammers succeed?

- ○ yes
- ○ no

Back Next Page 3 of 6

Academic Research: EFT/e-wallet SMS scams

If 'yes' to the previous answer

why did you think the scammers succeeded?

- ○ I was too busy, I did not verify the sender of the e-wallet or EFT deposit SMS e.g. +362626 for FNB
- ○ The number that sent the e-wallet or EFT deposit SMS was saved in my phone with the same name as the number used by my bank. e.g. +362626 for FNB
- ○ Other:

Back Next Page 4 of 6

Never submit passwords through Google Forms.

Academic Research: EFT/e-wallet SMS scams

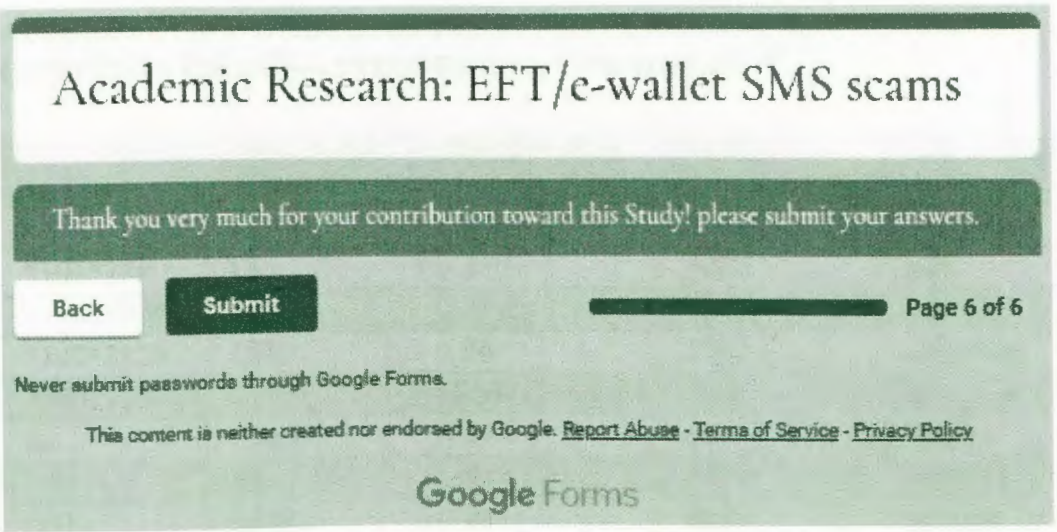
EFT or e-wallet scam detection app

If there is a smartphone app that automatically detects and warn users when they receive possible e-wallet or EFT scam SMSes, would you want to have it on your smartphone?

- ○ yes
- ○ no

Back Next Page 5 of 6

Never submit passwords through Google Forms.


Academic Research: EFT/e-wallet SMS scams
Thank you very much for your contribution toward this Study! please submit your answers.
Back
Submit
Page 6 of 6
Never submit passwords through Google Forms.
This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy
Google Forms

APPENDIX C - WORDS AND TERM FEATURES EXTRACTED FROM CONTENTS OF SMSes COMPRISING THE DATASET

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
***il	13h00	25	500gb	acidentlyy
*101*27#	15	250	5080	add
*140*295#	150	25sep	536534	address
*140*392#	1500	26	5989	addresses
*140*999#	15794	264813784384	5pm	adm
0	15oct	27	6	admin
00	16	299	600	advance
000	1648	2a	6014	adviser
01	16gb	2as1xd1zyde	6080	affordable
01345087678	16h40	2k19whp	69	after
02h00	16hrs	2nad	6kwh	again
03	17	2qqwe0h	750	agent
05	18	3	75320	ago
061	18201	30	7pm	ahead
0612992222	19	300	8	aino
07	1914	31	800	airtime
07h00	1974	3196	81	alarm
0811480480	1o1urhn6izjjsxs	32	823330	all
09	1tb	33648	845	almost
1	2	33652	850	already
10	20	35	8605	also
100	200	350	864483594	alt
1000	2000	3547	87490	am
102000	2010	362202	894830	an
1023	2018	37474	96	analog
104400	2019	38800	96684	and
10k	2019912032	3902	97765	answer
1189	2049	3903	9am	any
11th	21	3a	=	anymore
12	213627	3pm	@	anyone
123	22	400	a	anything
1250	2213	40176	aanâ¤kwh	anyways
125070	226	41	abaut	apologies
1251	23	45	able	app
128	23412	5	acc	application
12h	249	50	access	apply
13	24aug	500	account	appointment

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
april	bluevoucher	central	contract	disturbing
are	bonus	chance	convenience	dizzy
around	bookmarks	chances	convenient	do
arrived	boss	change	convert	documents
as	box	charge	cool	doeseb
asap	bra	charger	corner	doing
ask	brackets	chase	cos	don
asking	brand	chat	cotas	done
asleep	bring	cheaper	could	dont
assteria	bringing	check	coz	down
at	bro	chemist	coze	download
atm	broker	Chihingamu-yeva@gmail	credit	dramatically
attend	brother	choose	crest	dreams
auction	brur	church	cs	drive
aug	budget	claim	ctrl	drop
avail	busy	classes	cumng	due
available	but	clicking	cup	during
aweh	buy	client	current	early
axali	by	closed	customer	easywallet
babe	c	clouds	d344	ecblevy
back	cable	cmd	d751	eewa
balance	call	code	daily	efficiency
bank	called	collect	day	effortlessly
banking	cam	com	days	either
bar	came	combination	dc0pqc8dflh4-srasist0nr	electricity
be	cameras	come	deal	else
beautify	can	coming	dear	email
been	cant	commented	decision	emergency
before	car	company	declined	end
benefit	card	complaining	decoder	ends
best	cash	comprehensive	defcon	energy
bid	cashback	configured	del	enjoy
birthday	categories	confirm	dial	enter
bit	catering	confirmation	did	entrage
black	caused	connected	didn	entrance
blessings	cell	connection	diseertation	epl
blocked	cellphone	contact	dish	equipment

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
erase	fm	grysblok	hw	kaa
error	fnb	guarantees	i	kaukolelwa
esiku	for	guy	ibank	key
etc	forms	guys	id	keys
even	forward	gv8f47fb	ideal	kind
evening	forwarding	gw	if	kindly
event	free	had	im	kitaar
every	fresh	hai	important	know
everything	friday	hakahana	in	komatango
evng	fridges	happen	inagafutwa	konambango
ewalet	frie	happy	include	kosa
ewallet	from	hard	inconvenience	kowambo
exam	fsen****	has	increase	kristianh
examination	funds	have	information	kuume
exclusive	furniture	he	iniigwanithapo	laptop
expired	gakala	headache	install	laptops
expires	game	hear	installation	last
explorer	garage	help	installing	late
extended	gather	her	instalment	later
external	get	here	instead	launching
eyes	getting	hhfmfob3h-n7yheip9	insufficient	leaving
facebook	give	hi	insurance	lenovo
fantasy	gle	hike	interactive	leonard
far	go	him	interested	let
favor	going	hollard	internal	library
fb	gokulukadhi	home	invoice	licence
fbn	gold	hospital	invoices	lifestyle
feel	gone	hostel	is	like
fifa	good	hottest	issued	limit
final	goods	hours	it	link
find	gordon	house	its	linked
finder	got	household	join	little
fine	government	how	jules	ll
finish	gprs	hrs	july	lnb
finishing	graveyard	http	june	loaded
first	great	https	just	loan
flour	group	huawei	k	location

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
lock	monday	next	omzwa	passby
looking	mondjila	nga	on	password
loose	money	ngame	onandjokwe	pay
lot	monika	ngaye	once	paying
ltd	month	ngele	ondangwa	payment
lunch	months	ngeno	ondili	pc
ly	more	nkeloe	one	pelembe
lyalandula	morning	nmc	ongula	peni
m	moskopa	no	onkelo	people
magano	motor	nooma	online	perheps
mahangu	mounting	normal	only	phone
main	movenduka	not	open	physiotherapy
maintenance	movies	notebook	operating	pick
make	mtc	nov	opoto	pictures
makes	much	now	options	pin
mam	must	nsfaf	optout	pineas
man	mwadhin	number	or	pitty
may	my	nurses	organise	place
maybe	n	nust	otagapatwa	play
me	naaah	nâ¤	otandiyi	please
meet	nad	occupied	otawu	pls
meeting	nad400	occurred	other	plus
megumbo	nambahu	oct	others	pm
mejoyce	namboer	october	otjomuise	point
member	namcourier	of	otp	points
members	name	off	otwalalapo	policy
memorial	nampower	offer	our	possible
men	nawa	office	out	poto
messages	ndikutumine	offices	outer	ppl
min	ne	ok	ow	precious
mine	nee	oka	owutala	premiums
mini	need	okay	paid	press
minimum	needs	old	pamwe	price
mistake	neflevy	olefa	papers	prices
mms	nenge	olivia	part	prime
mobile	network	omeya	participate	print
model	new	omwalalapo	participation	printing

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
pritent	replying	series	spectacle	tax
private	repo	service	spend	taxi
probably	request	services	spree	team
proceed	reschedule	settings	ssc	tell
product	reserved	settled	stand	telling
professionals	respond	share	start	telongitha
proof	response	she	started	temp
provided	rest	shelf	statement	temporary
provider	retrieve	shem	station	tender
purchase	return	shilongo	stay	thameehileni
purchased	right	shitefa	std	than
queries	rimet	shop	stil	thank
quite	robert	shopping	still	thanks
quote	rocky	shot	stock	thanx
rain	root	should	stop	that
rates	rossing	shouls	straight	thats
reading	s	shuna	street	the
ready	s5	side	student	them
receipt	safe	sign	stuff	then
received	said	signings	style	there
reception	same	silver	submit	these
recieved	sample	since	submitted	thesis
reduce	sanlam	single	subscription	they
ref	satellite	sister	sun	thickener
reflect	saturdays	sized	sundays	think
refund	schedule	skip	sup	this
regards	school	sleeping	super	thnks
registration	sealed	smaller	support	though
reinstallation	secilia	sms	suppose	thought
reject	security	so	supposed	thursday
remember	see	some	sure	tiles
reminder	seeing	someone	survey	till
remote	select	son	t	time
rent	selling	soon	table	tires
renting	send	sorry	take	tme
replace	sent	sorted	talked	to
reply	sep	speak	tate	today

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
today@10	use	west	yet	1750
token	used	what	yo	17896
told	using	whatsapp	yoh	17945
tomorrow	utmost	when	yoku	1850
too	uusila	where	yomambo	1900
took	vaasa	whk	you	1950
touch	valid	who	your	19886
toward	verification	will	youtube	20155
town	very	win	zula	2150
transport	vid	windhoek	â	21878
tried	video	winning	*140*393#	22hrs
trophy	viral	wish	02	2300
true	voicemail	wishes	02306	23056
trust	voucher	with	04	23302
try	wait	withdraw	05592	2350
trying	waiting	withdrawals	06	24
ts	wallpapers	withdrawn	06129922	2400
tse	walvis	without	06204	2500
tuesday	want	wo	08	264818069225
turn	wanted	wola	09232	2750
tv	was	work	10098	2800
txt	wash	working	10987	28178
tyapa	watch	world	11	29
type	watsap	write	1100	2900
u	way	written	11809	29804
um	we	wrong	11hrs	3000
unam	wednesday	x	1300	30592
unavailable	week	xico	13486	3200
uncle	weekdays	y5	14	32563
understand	weekend	y7	1400	34
units	weekly	yah	15095	35996
up	welcome	yakunda	151610	36
upgrade	well	yeah	1600	362626
upgraded	welwitschia	years	16053	362626fnb
ur	went	yep	16533	3800
urgently	were	yes	16hours	39
us	wernhil	yesterday	1700	4000

Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature	Term or Word Feature
4150	76453	amakali	iiyambo	philip
43	76622	amon	j	pius
450	77749	amunyela	johannes	q
4500	780	amweutupa	karel	querries
50851	78036	angula	kaulinge	r
51	78166	augustus	kavau	shilunga
52	78659	b	l	smart
55	790	boas	lazarus	smit
56	86346	christoph	lewis	staden
56402	87012	cloete	matheus	thom
56749	900	coman	mbambo	toivo
56781	90180	d	mbango	uendji
57	90345	david	moongo	uupindi
580	90613	e	moses	v
59906	90923	eino	mouton	van
650	92782	erastus	mutika	victor
65433	930	f	ngonga	w
65611	950	g	o	willem
67652	95179	gumende	oiver	wyk
700	98001	gustav	p	y
70844	99918	h	paul	z
71296	aludhilu	iikombo	paulus	
76258	amaambo	iithete	petrus	