Application of Machine Learning Classification to Detect Fraudulent E-wallet Deposit Notification SMSes

Fillemon S. Enkono
Technologist, Department of Aeronautical Science, School of Military Science, University of Namibia
https://orcid.org/0000-0002-0891-4654

Nalina Suresh
Lecturer, Information Technology Department, School of Computing, University of Namibia
https://orcid.org/0000-0002-9846-7199

Abstract

Fraudulent e-wallet deposit notification SMSes designed to steal money and goods from m-banking users have become pervasive in Namibia. Motivated by an observed lack of mobile applications to protect users from such deceptions, this study evaluated the ability of machine learning to detect the fraudulent e-wallet deposit notification SMSes. The naïve Bayes (NB) and support vector machine (SVM) classifiers were trained to classify both ham (desired) SMSes and scam (fraudulent) e-wallet deposit notification SMSes. The performances of the two classifier models were then evaluated. The results revealed that the SVM classifier model could detect the fraudulent SMSes more efficiently than the NB classifier.

Keywords

m-banking, e-wallets, short message service messages (SMSes), deposit notification, fraud, ham SMSes, scam SMSes, detection, machine learning, classifiers, naïve Bayes (NB), support vector machine (SVM), classification accuracy (CA), feature extraction, feature selection

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

Recommended citation

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

1. Introduction and research problem

In the past decade, there has been significant growth of spam (unwanted messages) in email and short message service messages (SMSes), and a contemporaneous growth in the capabilities of mobile banking (m-banking) (Almeida et al., 2011; Shaikh & Karjaluoto, 2015). Among its capabilities, m-banking allows users to utilise mobile phones to make payments from their bank accounts to other users' electronic wallet (e-wallet) accounts. It also allows payment recipients to receive notification SMSes that acknowledge payments into their e-wallets. The widespread use of SMS notifications to acknowledge e-wallet deposits has served to establish such notifications as a trusted means of proof-of-payment for m-banking users. This trust has tended to make e-wallet recipients complacent, to the extent that they may neglect to verify the legitimacy of a payment upon receipt of a deposit notification SMS. This habit of non-verification has fostered the emergence of fraudulent SMSes that use false e-wallet deposit notifications in an attempt to deceive and defraud m-banking users (Arde, 2012; *Erongo*, 2016). It has been reported that m-banking users in Namibia, its neighbour South Africa, and other developing countries have suffered substantial losses, of both money and goods, through falling victim to fraudulent e-wallet deposit notification SMSes (Arde, 2012; Christopher & Kar, 2018; *Erongo*, 2016; Nagel, 2015).

This study was motivated by an observed lack of mobile applications for detecting fraudulent e-wallet deposit notification SMSes in order to safeguard m-banking users from the associated frauds. Our study, conducted in Namibia, evaluated the application of two machine learning classifiers to distinguish between ham (legitimate) SMSes and scam (fraudulent e-wallet deposit notification) SMSes. The classifiers tested were the naïve Bayes (NB) and support vector machine (SVM) models. The evaluation aimed to determine which of the two models could detect the fraudulent e-wallet deposit notification SMSes more efficiently. The ultimate aim was to establish which model would be a good candidate for implementation as an application for detecting fraudulent notifications on users' mobile devices.

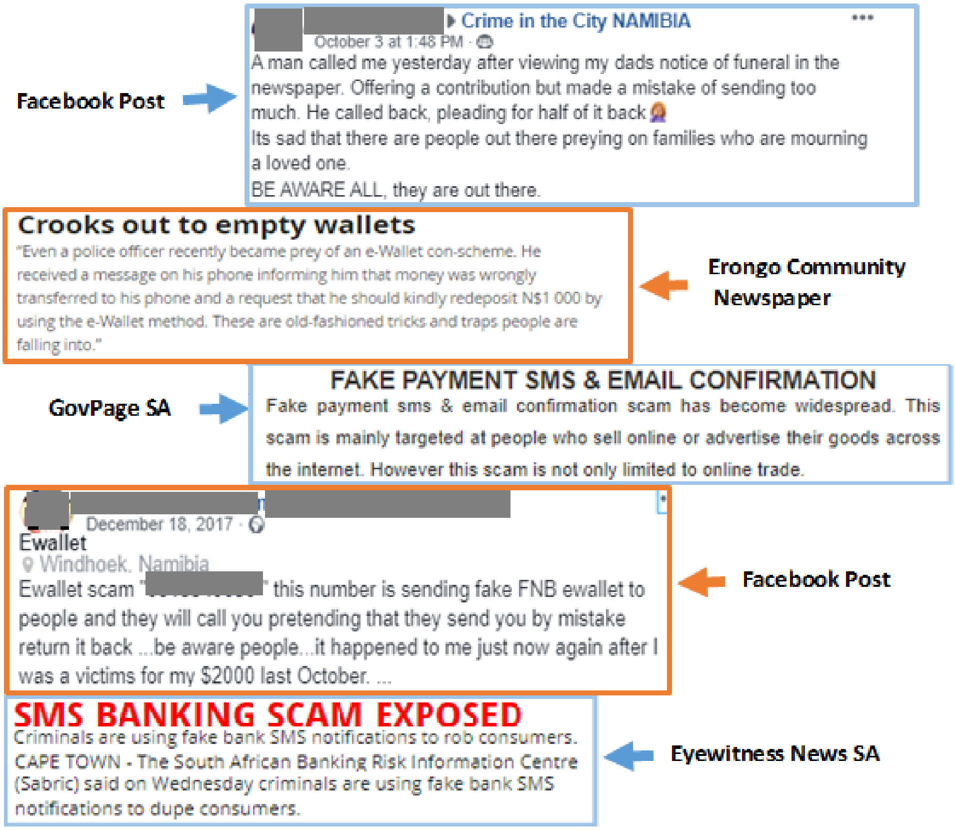
2. E-wallet deposit notification SMS fraud

In the typical scenario, the fraudster first obtains the m-banking user's mobile number from a source such as a website, a Facebook notice, or an advertisement. The fraudster then forges an e-wallet deposit notification SMS that purports to acknowledge the deposit of a certain amount and sends it to the m-banking user. The fraudster follows up with a call or SMS, and claims to have mistakenly deposited the amount specified in the bogus notification SMS into the user's e-wallet. The fraudster then asks the targeted user to refund the money via an e-wallet payment. The fraudster may seek to make the refund more palatable to the user by asking for only a portion of the amount supposedly deposited, and may also use other social engineering tricks to lure the user into falling for the fraud.

If the targeted user does not verify that the funds that the fraudster claims to have sent have actually been paid into their e-wallet, the user may fall for the fraud and make a payment from their account to the fraudster's e-wallet. In cases where the targeted user falls for the fraud, the fraudster simply withdraws the money and discards the SIM card used. Fraudsters use similar tricks in respect of goods, e.g., sending bogus e-wallet deposit notification SMSes to salespersons pretending to have paid for the goods, thus seeking to obtain goods without actually paying for them.

Since such frauds are relatively easy to detect and avoid under normal circumstances, fraudsters take advantage of specific situations in order to increase the target's vulnerability—e.g., communicating with someone who has had a death in the family, or with an online seller—as seen in Figure 1, which provides examples of fraud in Namibia and South Africa. Among the factors that create an enabling environment for e-wallet fraudsters are low SIM card costs and ready access to e-wallet services by anyone with a SIM.

Figure 1: Namibian and South African online postings on e-wallet payment notification fraud



[redacted] ▸ Crime in the City NAMIBIA
October 3 at 1:48 PM ·
A man called me yesterday after viewing my dads notice of funeral in the newspaper. Offering a contribution but made a mistake of sending too much. He called back, pleading for half of it back
Its sad that there are people out there preying on families who are mourning a loved one.
BE AWARE ALL, they are out there.

Crooks out to empty wallets

"Even a police officer recently became prey of an e-Wallet con-scheme. He received a message on his phone informing him that money was wrongly transferred to his phone and a request that he should kindly redeposit N$1 000 by using the e-Wallet method. These are old-fashioned tricks and traps people are falling into."

FAKE PAYMENT SMS & EMAIL CONFIRMATION

Fake payment sms & email confirmation scam has become widespread. This scam is mainly targeted at people who sell online or advertise their goods across the internet. However this scam is not only limited to online trade.

[redacted]
December 18, 2017 ·
Ewallet
Windhoek, Namibia
Ewallet scam "[redacted]" this number is sending fake FNB ewallet to people and they will call you pretending that they send you by mistake return it back ...be aware people...it happened to me just now again after I was a victims for my $2000 last October. ...

SMS BANKING SCAM EXPOSED

Criminals are using fake bank SMS notifications to rob consumers.
CAPE TOWN - The South African Banking Risk Information Centre (Sabric) said on Wednesday criminals are using fake bank SMS notifications to dupe consumers.

3. Literature review

SMS spam, smishing, and machine learning (ML) classification

Approaches to combatting spam email and SMSes generally involve the use of machine learning classification (see Akbari & Sajedi, 2015; Choudhary & Jain, 2017). Initial machine learning classification approaches to detecting SMS spam largely treated spam as a generalised set of undesired SMSes, without detection and delineation of specific types of spam (Abdulhamid et al., 2017). As malicious SMS phishing, a form of cyber-fraud known as "smishing", has become more prominent, approaches to spam detection have had to become more focused. The emergence of smishing has led to investigation of machine learning classifiers designed to specifically detect smishing SMSes (Goel & Jain, 2018; Jain & Gupta, 2018).

Further complicating the picture has been the emergence of legitimate SMSes with characteristics that overlap with those of spam and that could easily be erroneously treated as spam by existing spam detection or filtering systems (Reaves, Blue, Tian, Traynor, & Butler, 2016). These legitimate SMSes include those sent for purposes of advertising and promotion, and SMSes for verification codes and for password-reset codes, both of which users can receive from sender numbers that they could not have known beforehand (Reaves et al., 2016). This makes targeted approaches to detecting specific types of SMS spam even more necessary, especially for the types of spam, such as the e-wallet deposit notification frauds, that have the potential to cause users significant losses. The problem of fraudulent e-wallet deposit notification SMSes is relatively new, despite its reported pervasive manifestations, and this tends to explain why our literature survey did not manage to identify any work done on the application of machine learning classification to specifically detect fraudulent e-wallet deposit notification SMSes.

SMS datasets, feature extraction, and machine learning classification

Studying or employing machine learning classification of SMSes requires access to appropriate datasets of ham SMSes and spam SMSes (Abdulhamid et al., 2017). Methods often used to obtain the required dataset include obtaining SMSes from public databases (Ahmed et al., 2014), extracting SMSes from public web-based sources (Almeida et al., 2011), and collecting SMSes directly from users (Shahi & Yadav, 2014).

A study by Cormack (2008) explains that prior to applying machine learning to classify textual content, the content must first be represented as a collection of features derived from the text or from extrinsic information related to the text. Feature extraction is often employed to capture textual features for classifying texts, and the feature extraction process frequently produces multi-dimensional feature sets. It then becomes necessary to employ feature selection, so as to eliminate the features that are less significant in the classification of the text.

Various machine learning classifiers have been employed to classify SMSes for filtering or detecting spam. Some of the classifiers that are extensively used include naïve Bayes (NB), Random Forest (RF) and support vector machine (SVM) (Ahmed et al., 2014; Hedieh et al., 2016; Nagwani & Sharaff, 2017; Nuruzzaman et al., 2011). The NB classifier is widely used for SMS classification due to its simplicity and speed, while the common use of SVM and RF tends to be motivated by their high classification accuracy (CA)—often reported to be in ranges above 90% (Nagwani & Sharaff, 2017). For our study, we chose to test and compare the accuracy of the NB and SVM classifiers.

4. Methodology

The study employed an experimental research design, and used the Weka open source data mining software platform for the experiments.

The SMS dataset

We collected a dataset of 240 unique SMSes from Namibian m-banking users: 184 ham (i.e., normal and legitimate) SMSes and 56 scam (i.e., fraudulent) e-wallet deposit notification SMSes. The ham SMSes included legitimate e-wallet deposit notification SMSes. The ham SMSes, and some of the scam e-wallet deposit notification SMSes, were solicited from volunteers via invitations sent out on Facebook. The majority of the scam e-wallet deposit notification SMSes were extracted from user posts on public Facebook group (M-banking users habitually share examples of scam e-wallet deposit notification SMSes in online fora in order to warn other users). We then represented the 240 raw SMSes in terms of three attributes and a class specification, as outlined in Table 1.

Table 1: Attributes used to represent sample of raw SMSes

Attribute or Class	Description
senderNumLen	The length (total number of digits) of the mobile number of the SMS sender. (Banking institutions use short SMS codes to send legitimate e-wallet deposit notification SMSes, while fraudsters use normal (e.g., 10-digit) mobile numbers.)
content	The string of content in the body of the SMS.
contentLen	The length (number of characters) of the SMS body. Ham SMSes tend to have fewer characters than deposit notification SMSes (both legitimate and fraudulent).
smsClass	Whether the SMS is in the ham class or the scam e-wallet deposit notification class.

Feature extraction

We found that the three initial attributes used to define the raw SMSes—*senderNumLen*, *content* and *contentLen*—were insufficient to effectively classify the ham SMSes and the scam e-wallet deposit notification SMSes. We also found that the SMSes' contents contained numerous textual features that could be used to improve the classification. Hence, we employed Weka's StringToWordVector unsupervised filter in order to extract normalised word and term features from the SMSes' contents, for use, in addition to the *senderNumLen* and *contentLen* features, in classifying the SMSes.

Optimal classification features

A total of 1,223 features were extracted from the contents of the 240 SMSes in the dataset. With the addition of the *senderNumLen* and *contentLen* features, the set of features numbered 1,225. Weka's information gain (IG) feature selection algorithm was then used to select a subset of the 1,225 features that allowed optimal classification of the SMSes. The optimal classification features were determined by applying feature selection using different IG threshold (IG_{thshld}) values. Table 2 shows the number of features selected using different IG_{thshld} in the [0.0, 1.0] range, and the two classifier models' respective average CA for 10-fold cross-validation.

Table 2: IG_{thshld}, number of selected features, and the classifiers' average CA

IG_{thshld}	No. of selected features	Average CA	
		NB	SVM
0.000	119	0.9711	0.9876
0.025	48	0.9711	0.9917
0.050	25	0.9545	0.9628
0.075	22	0.9545	0.9628
0.100	21	0.9545	0.9628

Table 2 shows that a subset of 48 features, selected with IG_{thshld} = 0.025, allowed the NB and SVM models to classify the SMSes with optimal CA. These 48 features allowed the NB classifier model to maintain CA = 0.9711, its highest observed CA; and the SVM model to achieve CA = 0.9917, its highest recorded CA.

Table 3 shows the 48 word and term features that allowed the NB and SVM models to optimally classify the SMSes, along with their IG values, which indicate how much information each feature contributed towards the correct classification of SMSes that contain them.

Table 3: Features and their IG values

Feature	Feature IG value	Feature	Feature IG value	Feature	Feature IG value
e-wallet	0.6416	f	0.365	with	0.0368
na	0.6158	expires	0.3431	on	0.0368
dial	0.615	at	0.2779	your	0.035
*140*392#	0.6134	for	0.2682	of	0.035
sent	0.6028	you	0.132	will	0.035
contentLen	0.5421	362626	0.1281	can	0.0331
00	0.5392	the	0.0822	am	0.0331
select	0.5372	i	0.0686	me	0.0313
fnb	0.5309	valid	0.0626	*140*999#	0.029
expired	0.5233	it	0.0519	are	0.0277
press	0.5233	senderNumLen	0.0475	or	0.0277
proceed	0.5233	queries	0.0437	2350	0.026
services	0.5233	06129922	0.0437	1600	0.026
atm	0.498	and	0.0424	14	0.026
pin	0.498	to	0.0401	please	0.0259
new	0.4754	16hrs	0.0393	in	0.0259

Classifier models' training and evaluation

The dataset of the 240 SMSes, having been defined using the 48 optimal features, was then used for training and evaluating the NB and SVM classifier models. Supervised learning was employed, following a 10-fold cross-validation approach. The evaluation gave most weight to the models' capability to detect fraudulent e-wallet deposit notification SMSes, and was based on the following metrics adopted from works by Abdulhamid et al. (2017), Mahmoud and Mahfouz (2012), and Hedieh et al. (2016):

True positives (TP): The number of scam e-wallet deposit notification SMSes that are correctly classified.

True negatives (TN): The number of ham SMSes that are correctly classified.

False positives (FP): The number of ham SMSes that are falsely classified.

False negatives (FN): The number of scam e-wallet deposit notification SMSes that are falsely classified.

False positives rate (FPR): The rate of ham SMS misclassification.

$$FPR = \frac{FP}{FP+TN} FPR = \frac{FP}{FP+TN} \qquad \text{equation (1)}$$

False negatives rate (FNR): The rate of scam e-wallet deposit notification SMSes misclassification.

$$FNR = \frac{FN}{FN+TP} FNR = \frac{FN}{FN+TP} \qquad \text{equation (2)}$$

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

$$CA = \frac{TP+TN}{TP+FP+TN+FN} CA = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{equation (3)}$$

Precision: The proportion of SMSes classified as scam e-wallet deposit notifications that are correctly classified.

$$Precision = \frac{TP}{TP+FP} Precision = \frac{TP}{TP+FP} \qquad \text{equation (4)}$$

Recall: The proportion of the actual scam e-wallet deposit notification SMSes that are correctly classified.

$$Recall = \frac{TP}{TP+FN} Recall = \frac{TP}{TP+FN} \qquad \text{equation (5)}$$

F1-measure: The harmonic mean of precision and recall.

$$F1\ measure = \frac{2\ \text{x}\ Precision\ \text{x}\ Recall}{Precision+Recall} F1\ measure = \frac{2\ \text{x}\ Precision\ \text{x}\ Recall}{Precision+Recall} \text{equation (6)}$$

5. Comparative evaluation results

Table 4 presents the two classifier models' performances in terms of TP, TN, FP, FN, FPR and FNR evaluation metrics.

Table 4: Classification performance in terms of TP, TN, FP, FN, FPR and FNR

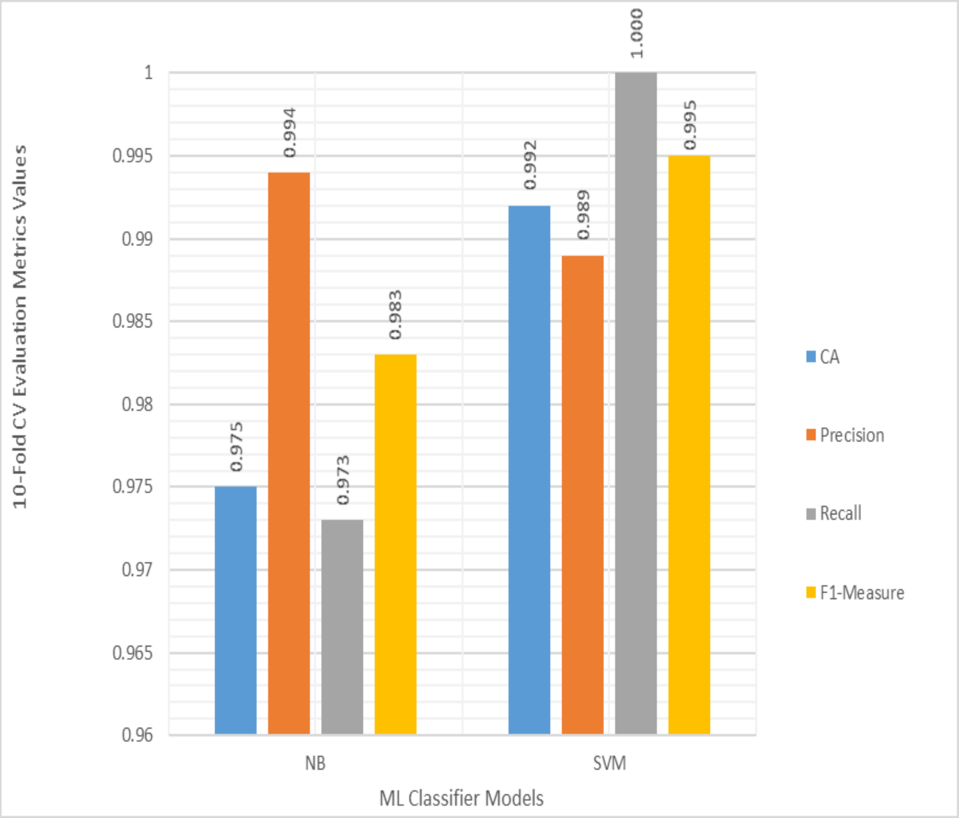
Metrics	10-fold cross-validation average results	
	NB	SVM
TP	17.700	18.200
TN	5.700	5.600
FP	0.100	0.200
FN	0.500	0.000
FPR	0.017	0.034
FNR	0.027	0.000

The results in Table 4 show that, on average, the SVM model correctly classified 18.2 scam e-wallet deposit notification SMSes (i.e., TP = 18.2) and misclassified 0.0 (i.e., FN = 0.0) compared to the NB's 17.7 and 0.5 respectively. This reveals that the SVM classifier model was more efficient than the NB model with respect to detecting scam e-wallet deposit notification SMSes. However, the results indicate the contrary about the models' capability to detect the ham SMSes, with the NB having correctly classified an average of 5.7 ham SMSes and misclassified 0.1, compared to the SVM's 5.6 and 0.2 respectively. The two models' FPR and FNR averages conform with the aforementioned results.

Because the emphasis of the study was on the models' capability to efficiently detect scam e-wallet deposit notification SMSes, we gave performance in terms of this criterion more weight than performance in respect of the detection of ham SMSes. Thus, the SVM classifier model was found to be superior to the NB model.

The graph in Figure 2 depicts the two classifier models' performance in terms of CA, precision, recall and F1-measure.

Figure 2: Classification performance in terms of CA, precision, recall and F1-measure



The results shown in Figure 2 indicate that the SVM model correctly classified more SMSes than the NB model, producing the highest CA. The two classifier models demonstrated contrasting performances in terms of precision and recall, with SVM achieving the highest recall while NB produced the highest precision. The differences between precision and recall, in terms of their respective equations (4) and (5) provided in the previous section, is represented by FP and FN. The two classifier models had contrasting FP and FN values due to their dissimilar performances with respect to correctly classifying the ham SMSes and the scam e-wallet deposit notification SMSes, as highlighted in the previous section. This caused the observed models' contrasting precision and recall values. The harmonic mean of precision and recall (i.e., F1-measure) helped to remove any ambiguity regarding which of the two models made a better overall classifier for both ham SMSes and scam e-wallet deposit notification SMSes, with the SVM model's F1-measure = 0.995 being superior to the NB model's F1-measure = 0.983.

6. Conclusions

In an attempt to contribute to solution of the problem of fraudulent e-wallet deposit notification SMSes, we trained NB and SVM classifier models to classify ham SMSes and scam e-wallet deposit notification SMSes, following which their performances were evaluated. The evaluation results indicate that the SVM classifier model can detect fraudulent e-wallet deposit notification SMSes more efficiently than the NB model. The SVM model's strength was highlighted by its FNR = 0.000, CA = 0.992, recall = 1.000 and F1-measure = 0.995, making it the more efficient model. Our envisaged future work is to extend this study by developing a mobile application or applications making use of the SVM classifier model to detect fraudulent e-wallet deposit notification SMSes on a user's mobile device.

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. https://doi.org/10.1109/ACCESS.2017.2666785

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. https://doi.org/10.7763/IJMLC.2014.V4.409

Akbari, F., & Sajedi, H. (2015). SMS spam detection using selected text features and boosting classifiers. In *2015 7th Conference on Information and Knowledge Technology (IKT)* (pp. 1–5). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/IKT.2015.7288782

Almeida, T. A., Gómez, J. M., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. In *2011 ACM Symposium on Document Engineering* (pp. 259–262). Association for Computing Machinery (ACM). https://doi.org/10.1145/2034691.2034742

Arde, A. (2012, January 8). EFT scammers fake their proof of payment to dupe you, the seller. *IOL*. Retrieved from https://www.iol.co.za/personal-finance/eft-scammers-fake-their-proof-of-payment-to-dupe-you-the-seller-1209136

Choudhary, N., & Jain, A. K. (2017). Towards filtering of SMS spam messages using machine learning based technique. In D. Singh, B. Raman, A. K. Luhach, & P. Lingras (Eds.), *Advanced Informatics for Computing Research: First International Conference, ICAICR 2017* (pp. 18–30). https://doi.org/10.1007/978-981-10-5780-9_2

Christopher, N., & Kar, S. (April 24, 2018). Mobile wallet scam: As next-gen spenders go cashless, e-wallet scamsters too are getting creative. *The Economic Times*. Retrieved from 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. https://doi.org/10.1561/1500000006

Crime in the City NAMIBIA (n.d.). [Facebook group]. Retrieved from https://www.facebook.com/groups/616819348371856/permalink/2462372400483199/

Erongo. (2016, December 1). Crooks out to empty wallets. Retrieved from http://www.erongo.com.na/news/crooks-out-to-empty-wallets/?

FNB Namibia Classic Clashes. (n.d.). [Facebook group]. Retrieved from https://www.facebook.com/FNBNamClassicClashes

Goel, D., & Jain, A. K. (2018). Mobile phishing attacks and defence mechanisms: State of art and open research challenges. *Computers & Security*, *73*, 519–544. https://doi.org/10.1016/J.COSE.2017.12.006

Govpage. (2017). Fake payment SMS & email confirmation. Retrieved from https://www.govpage.co.za/general-scams-warnings/fake-payment-sms-email-confirmation

Hedieh, S., Parast, G. Z., & Akbari, F. (2016). SMS spam filtering using machine learning techniques: A survey. *Machine Learning Research*, *1*(1), 1–14. https://doi.org/10.11648/j.mlr.20160101.11

IOL. (2019, March 25). More than R12.8bn sent from eWallet transactions in six months. Retrieved from https://www.iol.co.za/business-report/economy/more-than-r128bn-sent-from-ewallet-transactions-in-six-months-20071814

Jain, A. K., & Gupta, B. B. (2018). Rule-based framework for detection of smishing messages in mobile environment. *Procedia Computer Science*, *125*, 617–623. https://doi.org/10.1016/J.PROCS.2017.12.079

Mahmoud, T., & Mahfouz, A. (2012). SMS spam filtering technique based on artificial immune system. *International Journal of Computer Science Issues*, *9*(2), 589–597.

Nagel, E. (2015, June 5). Watch out for these common payment scams [Blog post]. *Gumtree*. Retrieved from https://blog.gumtree.co.za/watch-out-for-these-common-payment-scams/

Nagwani, N. K., & Sharaff, A. (2017). SMS spam filtering and thread identification using bi-level text classification and clustering techniques. *Journal of Information Science*, *43*(1), 75–87. https://doi.org/10.1177/0165551515616310

Nuruzzaman, M. T., Lee, C., & Choi, D. (2011). Independent and personal SMS spam filtering. In *2011 IEEE 11th International Conference on Computer and Information Technology* (pp. 429–435). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/CIT.2011.23

Reaves, B., Blue, L., Tian, D., Traynor, P., & Butler, K. R. B. (2016). Detecting SMS spam in the age of legitimate bulk messaging. In *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks - WiSec '16* (pp. 165–170). Association for Computing Machinery (ACM). https://doi.org/10.1145/2939918.2939937

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. https://doi.org/10.4236/ijis.2014.41004

Shaikh, A. A., & Karjaluoto, H. (2015). Mobile banking adoption: A literature review. *Telematics and Informatics*, *32*(1), 129–142. https://doi.org/10.1016/J.TELE.2014.05.003

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