Machine Learning Sms Spam Detection Model

Andrew Kipkebut[1], Moses Thiga[2] Elizabeth Okumu[3]
School of Science , Engineering and Technology
Kabarak University, P.O. Box Private Bag, Kabarak, 20157, Kenya
Tel: +254 0719499615, Email: akipkebut@kabarak.ac.ke

Abstract

Millions of shillings are lost by mobile phone users every year in Kenya due to SMs Spam, a social engineering skill attempting to obtain sensitive information such as passwords, Personal identification numbers and other details by masquerading as a trustworthy entity in an electronic commerce. The design of efficient fraud detection algorithm and techniques is key to reducing these losses. Fraud detection using machine learning is a new approach of detecting fraud especially in Mobile commerce. The design of fraud detection techniques in a mobile platform is challenging due to the non-stationary distribution of the data. Most machine learning techniques especially in SMs Spam deal with one language. It is in this background that the study will focus on a client side SMs Spam detection in Kenya's mobile using machine learning. Naive's Bayes algorithm was used for this purpose because it is highly scalable in text classification. The contributors of Corpus include mobile service providers in Kenya and selected mobile phone users. Machine learning and data mining experiments were conducted using WEKA .The results and discussions are presented in form of descriptive statistics and detection metrics, the model registered an overall classification accuracy of 96.1039% .

Keywords: Algorithm, Classification, Detection, Machine learning, Naïve bayes, WEKA.

1.0 Introduction

The goal of machine learning is to improve the performance of a computer program with experience. There are a lot of tasks that can be solved using machine learning, including speech recognition, playing games, automatic driving of a vehicles, anomaly detection medical diagnosis and data mining (Sinclair, C., Pierce, L., & Matzner, S. (1999). A range of algorithms have been invented for machine learning, that uses the fields of artificial intelligence, probability, statistics, information theory, neurobiology and others. Some of these algorithms include decision trees, Support Vector machines (SVM), Artificial Neural Networks (Almomani, et al., (2013)), Naïve Bayes algorithm, Decision trees among others. Machine learning is a fascinating field of artificial intelligence where investigation on how computer agents can improve their perception, cognition, and action with experience. Mansfield Devine (2017) defined SMS phishing (SMS Spam) as a form of criminal activity that uses social engineering techniques in an attempt to harvest credentials such as passwords, PINs (personal identification numbers) and other details by masquerading as a trustworthy entity in an electronic communication using Short Message Service (SMS)

1.1 Statement of the problem

SMS spam is real and a growing problem largely due to the availability of very cheap bulk pre-pay SMS packages and the fact that SMS stimulate higher response rates as it is a trusted and a personal service. The Short Messaging Service (SMS) mobile communication system is attractive for criminal gangs for a number of reasons i.e. it is easy to use, fast , reliable and affordable technology (Delany S. J , Buckley M,& Greene D ,2012). The presence of lack of a unifying model is perceived as a hindrance to the further development of the field of machine learning especially in Sms spam detection. Many approaches proposed, regardless of their effectiveness, focus on a specific aspect or language and most of them do not have integrated approach and are not exhaustive.

1.3 Main objective

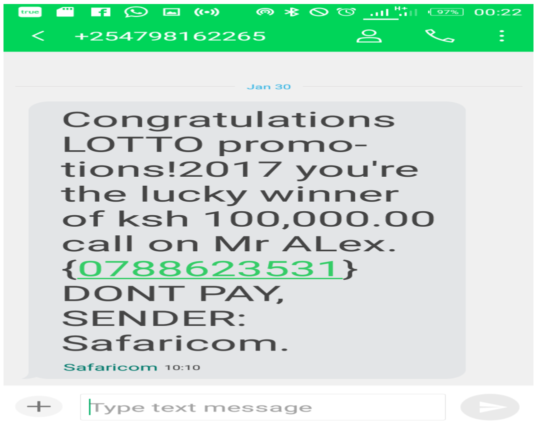
The main objective of this research is to evaluate a machine learning Sms Spam detection model.

1.4 Specific objectives

- To develop Spam detection model that can be used to detect Spam messages in Kenya .
- Demonstrate the use of machine learning in classifying messages as either Spam or not.
- To test the machine learning model through the use of a prototype.

2.0 Literature Review

Mobile phones have completely changed the way people communicate and interact. You can call, send text messages, read emails, play games as well as read and edit documents on the fly. Today, the mobile phone has become part and parcel of many people's lives. Leaving the house without your cell phone is like leaving your brain at home, some people may not function without the phone. According to Joo J.W, Moon S.Y, & Singh S, (2017) SMs Spam has continued to grow and evolve in popularity as a social engineering tool for cybercriminals. SMs Spam can trick the user into clicking on a link in a text message which can lead the user to entering personal data. The objective is to gain access to sensitive information like usernames and passwords. Additionally, many SMs Spam messages will include links with malware waiting on the other side for anyone who clicks on them. If you click on the infected web site link , it may download malware, which compromises your mobile device or the web site will ask you to input personal information such as, social security number, credit card type, credit card number and PIN etc. If you call the phone number given, it will sound very official and will ask you to input personal information such as, Identification number, credit card type ,digital wallet pin among others. An SMs Spam text message is determined on the basis of the basic attribute of the text message. Whether the text message includes a Universal Resource Locator(URL) or a telephone number or just plain sentences (Kang, S. & Kim, S., 2013) as shown in figure 1.



Congratulations LOTTO promotions!2017 you're the lucky winner of ksh 100,000.00 call on Mr ALex. {0788623531} DONT PAY, SENDER: Safaricom.

Safaricom 10:10

Figure 1: An SMs Spam example (source, Researcher).

2.1 Naïve Bayes classifier

As stated by Russell Stuart & Norvig Peter (2016) Naive Bayes classifiers are a family of simple probabilistic classifiers based on Bayes' theorem with strong independence assumptions between the features. It has been studied extensively since the 1950s., and was introduced under a different name into the text retrieval community in the early 1960s, it remains a baseline method for text categorization, the problem of judging messages as belonging to one category or the other (such as spam or legitimate) with word frequencies is used as features. With appropriate pre-processing, it is competitive in the classification domain with more advanced methods including support vector machines (Rennie J, Shih L ,Teevan J,& Karger D, 2003).

Naive Bayes (NB) is a classifying algorithm as shown in figure 2 uses data about prior events to estimate the probability of future events. Typically it's best applied to problems in which the information from numerous attributes should be considered simultaneously in order to estimate the Probability of an outcome. While many algorithm typically ignore features with weak effects, this technique uses all available info to subtly change predictions (Raghav Bali, Dipanjan Sarkar ,& Brett Lantz ,2013). Most of the current spam email detection systems use keywords to detect spam emails. These keywords can be written as misspellings eg: baank or bannk instead of bank. Misspellings are changed from time to time and hence spam email detection system needs to constantly update the blacklist to detect spam emails containing misspellings (Renuka & Hamsapriya,2010).A Naive Bayes classifier will converge quicker than other models, it requires less training, it is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Although the independence assumption may seem sometimes unreasonable, its performance is usually reasonably good, even for those cases (Romero, 2010).It is called Naïve classifier because its assumes independent features.

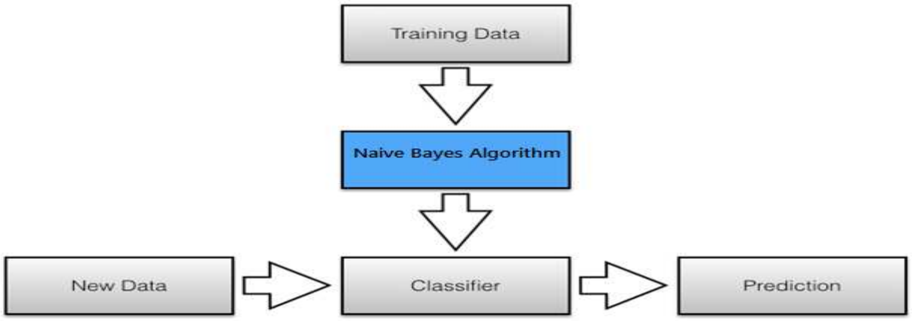


Figure 2 Naïve classifier approach (Sebastian Raschka, 2014)

.2.2 Conceptual Design

Yadav et al.,(2011) provides an approach is similar to Deng & Peng (2006) in that they propose a client side Naive Bayes filter which uses the occurrence of keywords that appear in spam messages to determine a spam score. Messages that score above a certain threshold are labelled as spam. Their solution also requires user feedback to confirm and correct errors made by the classifier and therefore their filter can learn new spam keywords from client reports to a central server, which are in turn pushed out to other clients. The research will be informed by the conceptual design in figure 3. The data is pre-processed which includes formatting of the data, then trained using Naïve Bayes ML algorithm after which its tested and lastly classified to determine which class they belong to.

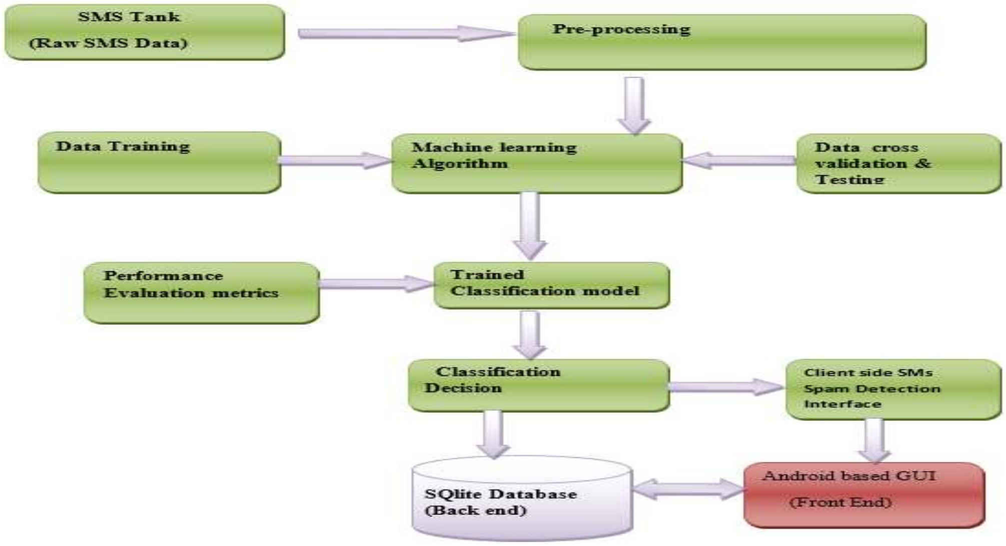

SMS Tank
(Raw SMS Data)
Pre-processing
Data Training
Machine learning Algorithm
Data cross validation & Testing
Performance Evaluation metrics
Trained Classification model
Classification Decision
Client side SMs Spam Detection Interface
SQlite Database (Back end)
Android based GUI
(Front End)

Figure 3 : Conceptual Design   Source Researcher

3.0 Methodology

This chapter provides a description of the approaches that were adopted in carrying out the research . Naive Bayes algorithm that uses Bayes theorem as shown in formula below was used for detecting whether a message is Spam or not.

$$P(spam|word) = \frac{\mathrm{P(spam).P(word|spam)}}{\mathrm{P(spam).P(word|spam) + P(non - spam).P(word|non - spam)}}$$

Several measurement methods were typically used for comparing results of classification Some of these methods included precision, recall, accuracy, true positive, false negative, true negative and false negative-rates (B.K. Bharadwaj & S. Pal, 2011).Machine learning experiments were employed in this research using WEKA. WEKA (Waikato Environment for knowledge analysis) an Open source data mining tool written in Java programming that can be used for collection of machine learning algorithms data, data pre-processing, classification, regression , clustering and visualization. It contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces (Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, & Witten I. H, 2009).

3.1 Prototype development.

The SMs Spam detection prototype development involves an iterative SDLC (Software development life cycle) , a process of dividing software development into distinct steps that contains finite activities , the steps include ,Problem definition, Data collection , Data preparation, spot check on algorithm, Training of the model ,Evaluation of model performance , data visualization and prototype implementation. On these activities there is a lot of flexibility. The greatest thing in using automated tools is that you can always go back a few steps (iteration) and insert a new transform of the dataset and re-run experiments in the intervening steps to see what interesting results come out and how they compare to the experiments executed before. The algorithm was trained on the training dataset and evaluated against the test set. This involves selecting a random split of data (66% for training, 34% for testing).

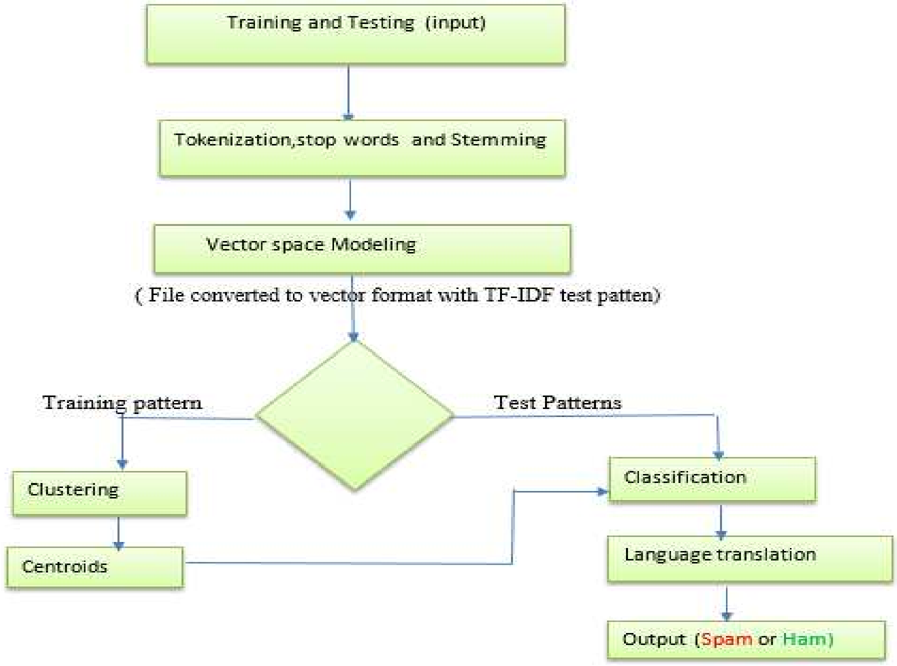


Fig 4 Flow chart of the proposed prototype

The flow chart in Fig 4 involves training and testing of the data , Tokenization , stop words and stemming, Vector space modelling using TF-IDF transformation. This will result into a training pattern and a test pattern that will be used for classification. Clustering was also done in order to group the messages as either spam or Not , this was done using K – means Clustering algorithm.

4.0 Results Findings and Discussions

This chapter gives an overview of the results, findings and discussions .An experiment was done to examine 1001 SMS as Spam or not a Spam. This analysis was done with reference to the three objectives aforementioned in the research objectives. These findings were used to explain the results and future work. WEKA tool was used to read stored SMS data from a file and was further structured for the learning algorithm interpretation.

4.1 Data Testing

A test data (Artff file) containing 322 instances (1/3 of total 1001 instances) with missing class(?) was used to test the full training data. Stemming was done using potter stemmer algorithm, This generated 1115 keyword based attributes that include Cash, safaricom ,win, bank among others,the Table 1 gives a the summary of keywords which includes its mean, standard deviation weighted sum and precision for each class.The naives bayes algorithm correctly classified 962 instances and incorrectly classified 39 of the 1001 instances which gives a percentage of 96.1039% and 3.8961% respectively. It leads to an accuracy of 96.1039% as shown in Table 2 and Table 3, from the confusion matrix the naïve bayes r learner was able to get 207 +755 correct classifications, and made 22+17 mistakes , this is fairly good compared to the nature of the algorithm.

Table 1 Some keyword -attribute statistics

Attribute	**Class**	
safaricom	**Spam(0.22)**	**Ham (0.78)**
Mean	0.0214	0
Std deviation	0.627	0.5327
Weight Sum	224	777
Precision	3.1928	3.1928

Table 2 Evaluation on training Set

Correctly classified instances	962	96.1069%
Incorrectly classified instances	39	3.8961%
Kappa statistic	0.8887	
Mean absolute error	0.0439	
Root Mean Squared error	0.1769	
Relative absolute error	12.6162%	
Root relative squared error	42.4404%	
Total number of instances	1001	

Table 3 Detailed accuracy by Class

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC**	**PRC-Area**	**Class**
Weighted Avg	0.924	0.028	0.904	0.924	0.914	0.889	0.989	0.973	**spam**
	0.972	0.076	0.978	0.974	0.975	0.889	0.989	0.997	**ham**
	0.971	0.065	**0.961**	0.961	**0.961**	0.889	0.989	**0.991**	

4.2 Conclusion

SMS spam filtering is an important issue in mobile commerce security and machine learning techniques; The quality of performance Naıve Bayes classifier is also based on datasets used. In this thesis Naive Bayes classifier has shown highest precision in Sms Spam detection. By looking at the words that are present within the message, the classifier was able to correctly classify the message as either Spam or Not. Using a model such as these mobile users can detect Spam messages using their phones therefore reducing fraud.

This model can be improved by looking at the messages that were mis-classified and understanding why this happened.

4.3 Recommendations

i) Some WEKA Visualizations features were not very clear e.g. the J48 visualization tree, this is because the tree was too large, other commercial software such as scikit-learn, RapidMiner may be used for this purpose.

ii) To avoid loss of money through SMs Spam, the government of Kenya needs to provide user training/education on Social engineering attacks especially on mobile phone.

iii) To speed up the training and testing , the researcher recommends thorough preprocessing of data. Use of a computer at least of 1.10 GHZ , 4GB RAM or above especially when handling large data is highly recommended

iv) Profiling of frequent sms Spam phone number may also help to curb this menace

v) Client side detection may not be enough. An adoption of a server side detection mechanism from the service providers such as Safaricom, Airtel and Telkom will help to reduce the damage of SMs Spam .

4.4 Areas of Further research

This study focussed on SMs Spam detection using naive bayes machine learning algorithm. However there are almost infinite text classification ways to detect SMS Spam that needs to be researched on. In this study the accuracy of the model can be improved with considering large data set and restrict the algorithm model to ignore normal dictionary words and instead use frequently used spam words in any language including Slang language (ghetto language) and local vernacular languages.

References

A Almomani, BB Gupta, T Wan, A Altaher (2013) Phishing Dynamic Evolving Neural Fuzzy Framework for Online Detection Zero-Day Phishing Email. Indian J. Sci. Technol. 6, no. 1, 3960–3964 .

A- Kwee, et al (2009), "sentence-Level Novelty Detection in English and Mafay." in advances in knowledge and discovery and Data mining vol. 5476. T. Theeramunkong. et al. Eds., cd: Springer Berlin Heidelberg., pp. 40-51.

B..Blankertz, G. Dornhege, C. Schafer, R. Krepki, J. Kohlmorgen, K.-R. Muller, V. Kunz-mann,

Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: Methods and data. *Expert Systems with Applications*, *39*(10), 9899-9908.

Mansfield-Devine, S. (2017). Bad behaviour: exploiting human weaknesses. *Computer Fraud & Security*, *2017*(1), 17-20.

Kang, A., Lee, J. D., Kang, W. M., Barolli, L., & Park, J. H. (2014). Security considerations for smart phone smishing attacks. In *Advances in Computer Science and its Applications* (pp. 467-473). Springer, Berlin, Heidelberg.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 616-623).

Sinclair, C., Pierce, L., & Matzner, S. (1999, December). An application of machine learning to network intrusion detection. In *Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99)* (pp. 371-377). IEEE.

Wang, C., Zhang, Y., Chen, X., Liu, Z., Shi, L., Chen, G., Qiu, F., Ying, C., & Lu, W. (2010).A behavior-based SMS antispam system. IBM Journal of Research and Development , 54 , 3:1{3:16.

Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., & Naik, V. (2011, March). SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications* (pp. 1-6).