Case Based Architecture to Categorize SMS in Mobile System

SMS Categorization for Future Security and User Convenience

[1]Neetu Gupta, [2]Ms. Sonal Arora

[1]Student, [2]Assistant Professor,
[1] Department of Computer Science and Engineering,
[1]DPGITM (MDU University), Gurgaon, India.

***Abstract**:* The intense growth of smart mobile phones and users has contributed to the growth of online or offline Instant Messaging and SMS usage as an alternative way of Transaction and communication. Along with the faith they instinctively have in their devices, makes this kind of messages a congenial environment for spammers. In fact, reports distinctly shows that volume of spam over Instant Messaging and SMS is rapidly increasing year by year. This represents a challenging problem for classical filtering methods these days. Smishing this term represents a phishing in SMS/Messages called as SMS-phishing is a cyber-security attack, which utilizes Short Message Service (SMS) to steal personal data/credentials of mobile users. The faith level of mobile users on their smartphones has attracted attackers to perform various mobile security attacks like SMS-Phishing. In this paper, we implement the SMS-Case-based data mining classification approach to classify them subpart of SMS category by detecting of Illegitimate/Smishing messages. This proposed approach identified more than fifteen Cases which can efficiently classify them into three parts like- Primary/Useful, Illegitimate/Fraud/Smishing SMS. Furthermore, our approach applies different classification algorithms to train these renowned Cases. Since the SMS/ Messages text are usually short and generally written in Lingo language, we have used text normalization to convert them into standard form to get better Cases [1]. The performance of the proposed approach achieved more than 99% true negative rate. Further, the proposed approach is very effective for the detection of the zero hour attack too and we will classify them into three subparts, like Primary/Useful, Illegitimate/Smishing, Promotional/Others. This classification data can be used to make a better enhanced existing SMS/Message application, by providing an enhanced version of SMS application

Keywords:
SMS, Message Analysis, Smishing, Illegitimate, SMS Classification, Data Mining, Cases, Phishing, Short Messages,

Benefits:
Using this proposed solution giant organization can enhance their existing mobile Message/SMS application to the user security and convenience.

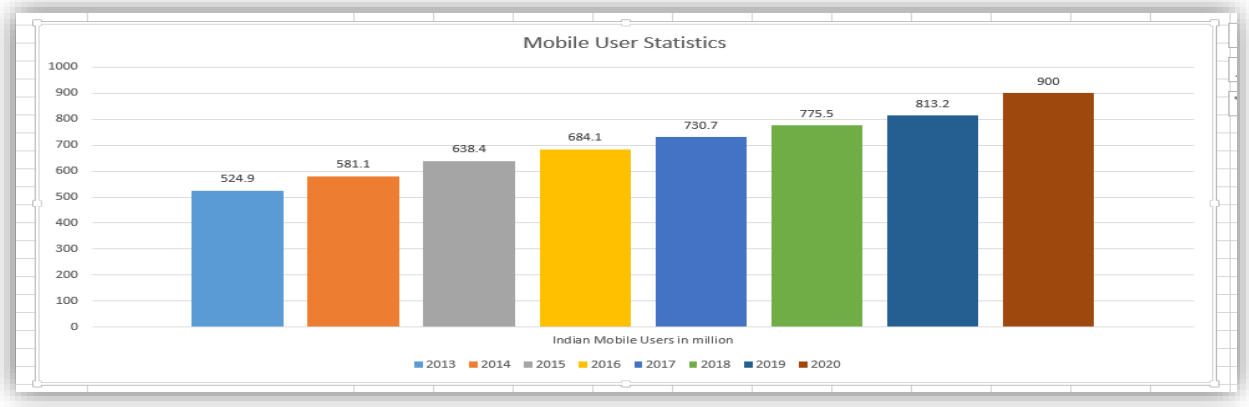
Why this approach/ need of this solution:

As SMS/Instant messages has become alternative way of transaction and communication, so there is a need to make SMS services more secure and classified.

I. INTRODUCTION

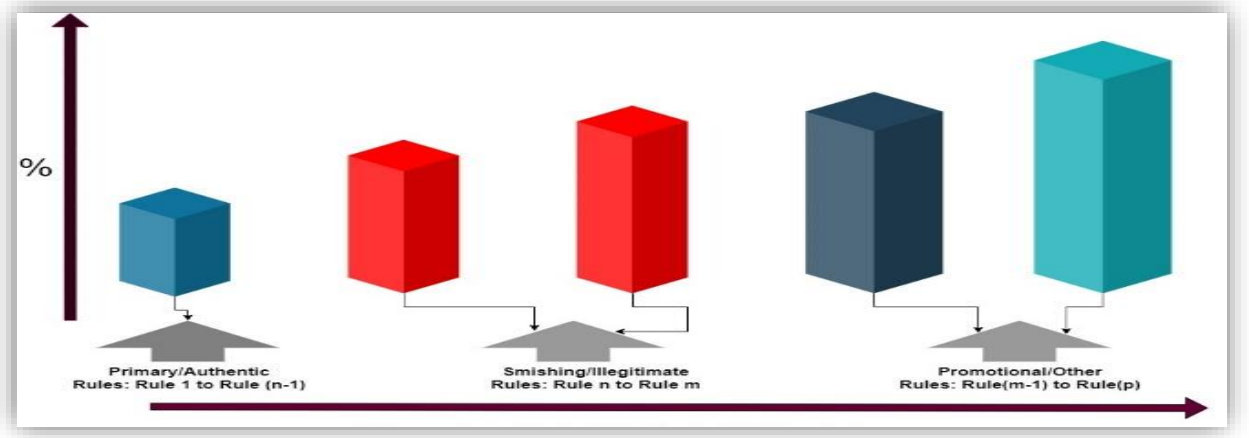
Short text messaging has become a mean of communication for an intense peoples these days. IM/SMS are clearly the leading way of communication. In fact, estimation says that near about 1.5 billion messages are sent a day by considering just SMS. The popularity of (SMS) has been growing over the last decade. The intense growth of smart mobile phones has contributed to the growth of online or offline SMS usage as an alternative way of Transaction and communication. For businesses text messages are simpler than even emails this is often because while 96% of mobile users read their sms by the top of the day about 85% of the emails remain unopened sms comparison 2018 [3-4]. Hence it's easy to know why sms has grown into a multibillion dollar commercial industry unfortunately over the years smartphones have also become the target referred to as short messages spam. SMS Spam refers to any worthless/irrelevant text messages delivered using mobile networks. They are strongly annoying to users [3, 4]. Example of a fraud/fake spam text message is like ''Hurray!! CONGRATULATIONS!!: YOUR MOBILE NO HAVE WON 1000,000 IN YOUR ACCOUNT— MOBILE DRAW USA,TO CLAIM PRIZE SEND BANK DETAILS, NAME, AGE, MOBILE NO,ADDRESS, TO--'. This kind of SMS came not from only national but International also. A survey identified that 69% of mobile phone users are affected by SMS Spam, with teenagers being the worst affected community [1].

Fig1: Mobile Phone User analysis in India [2]



Now a day's mobile security may be a major concern because attackers have diverted their mind from Computers to smartphones due to technology growth. Moreover, people are more attracted towards smartphones because it may be a portable and multi-functioning device, Smartphones are more popular now a days as compared to laptops due to their small screen size, lower cost , and portability. consistent with Dimensional Enterprise Mobile security Survey report and it shows that Smishing attack stands at the second position altogether quite mobile devices attacks [14]. There's two sort of security methods are wont to identify Illegitimate/fake mobile SMS. The primary method is that the blacklist based technique that forestalls the incoming SMS from the fake sources [17]. However, blacklist-based techniques don't cover all the fake sources, as an attacker can buy any mobile number to send the Illegitimate/fake/bogus SMS. The second sort of solution is predicated on the machine learning algorithm during which various features are extracted and compute from the SMS to require appropriate decision. The advantage of the machine learning based technique is that it can detect the fake message coming from any source. Data processing methods help within the feature extraction and finding the relation between them [16]. These approaches identifying hidden knowledge from datasets in terms of Cases and make the choice supported extracted Cases. Human easily understands these Cases and their Cases are written within the sort of IF condition THEN action. In this paper, we employed the Case-based data processing classification approach within the prediction of useful/illegitimate/promotional/offers SMS. We've used WEKA tool to classify SMS/Messages for data processing. We study the varied characteristics of text messages thorough then found fifteen Cases which may efficiently classify SMS to the subcategory. We then use Case-based classification algorithm namely Bayes Net, SMO, Bayesian classifier, Updated Bayesian Classifier, Decision Tree, ZeroR, OneR, and lots of more to use these Cases. In this, we've also identified the illegitimate/Smishing messages. Moreover, we recognize the simplest Case-based classification algorithm within the classification of Smishing messages. The performance of the proposed approach is evaluated, and it achieved quite 96% of true negative rate and 99% true positive rate.

Fig 2: SMS frequency analysis



(Primary/Useful, Illegitimate/Smishing, Other/ Promotional/Offers)

2. Review Of literature

Over the years, data scientists have proposed several ML models to spot Spam and not Spam. These aren't only for mobile text messages but also email spam and on social network platforms like facebook twitter delany et al [23] provided a survey of existing works for filtering spam sms. They mostly covered articles that relied on traditional machine learning approaches but not deep learning for instance, [24] compared Bayesian classifier with other classification algorithms and located that the previous was better to classify Spam text messages. Androulidakis et al. [25] proposed another model to filter Spam messages. Their model was supported the Android OS during which the users mobile control was wont to filter the Spam. The model checked the knowledge of message senders against a previously defined spammer list so when a message came from the users present within the list of spammers it had been treated as spam else not spam zainal et al.

Related Work

This section discusses the various existing mobile Text SMS classification detection techniques. The existing mobile classification and detection techniques divide into following section.

a) User Knowledge/Education Based Scheme

The educational based solutions focuses on educating the mobile users about the characteristics of phishing message through training, workshop and awareness programs so that they correctly identify the phishing attack [8]. However, the phishing attack becomes successful due to human flaws and ignorance. This conceptual knowledge may help the users in avoiding phishing attacks.

b) Technical solutions to mitigate mobile phishing attack

The technical solutions are cost-effective and straightforward to implement as compare to educational based solutions. In this, Amrutkar et al. [9] suggested mechanism named KAYO, which differentiates between the malicious and genuine mobile webpages. It detects mobile malicious pages by measuring 44 mobile features from webpages. Among 44 features, 11 are newly identified mobile specific features. KAYO's 44 feature set is split into four classes namely HTML, mobile specific, URL and JavaScript features. Joo et al. [6] proposed a model 'S-Detector' for detecting Smishing attack. They used Naïve Bayesian Classifier in their system to filter Smishing messages by finding the words used more often in these messages. S-Detector consists of SMS monitor, SMS analyzer, SMS determinant, and Database. Foozy et al. [7] proposed a Rule-based methodology to filter Illegitimate/Smishing messages from spam messages. Authors applied two Rule namely 'winner announcement' and 'marketing advertisement'. They need applied the Bayesian technique in WEKA tool to see the accuracy of Smishing, spam and ham messages. Alfy et al. [15] proposed a spam filtering model for both email and SMS. The proposed technique used 11 features namely presence of URLs, likely spam words, emotion symbols, special characters, gappy words, message metadata, JavaScript code, function words, recipient address, discipline and spam domain. they need evaluated their proposed model with five email and SMS datasets. Within the literature, we will conclude that no single technique exists which will detect illegitimate/Smishing attacks efficiently. Therefore, we'd like a way which will protect the user against Illegitimate/fake/Smishing attacks.

3 Research Methodology

In this we have discuss our proposed methodology of classifying messages into different categories by using some cases and accordingly we will classify them using WEKA tool and will identify the accuracy, further will also use some detection cases to detect illegitimate/Smishing SMS/ Messages.

The proposed approach is a model to filter SMS, protects the user from the phishing SMSs by blocking these messages and also implemented system can classify them into different category and delivering only Normal ones to the mobile user instead of making all into a single category it will further filtered into different category like- Primary/Useful, illegitimate/fake, Other/Promotional. The SMS detection is a type of ternary classification problem where a message can be the divide in three categories (i.e., Primary, Illegitimate/fake, Other/Promotional). Illegitimate message is a dangerous spam message that steals personal data/credentials. As per our research and observation, we find the followings characteristics of fraud message:

- It can have .exe message content in the link form.
- Now a days a different format seen like SMS includes.txt files.
- It can have any honey coated audio/video content that can trap the user.
- Advertising for offers/ Promotions.
- It can have the bogus fake links. Advertising something like providing free minutes, etc.
- It van have email address or a phone number.
- Links can have harmful viruses.
- Machine Recorded voice/Self-answering SMS asking the user to subscribe or unsubscribe any service.
- Announcing to users as a winner for fake contest and attract him using the prize money.
- Intended to spread some fake news..
- Message can have link and link have steganography.
- Long SMS can have fake. Etc.

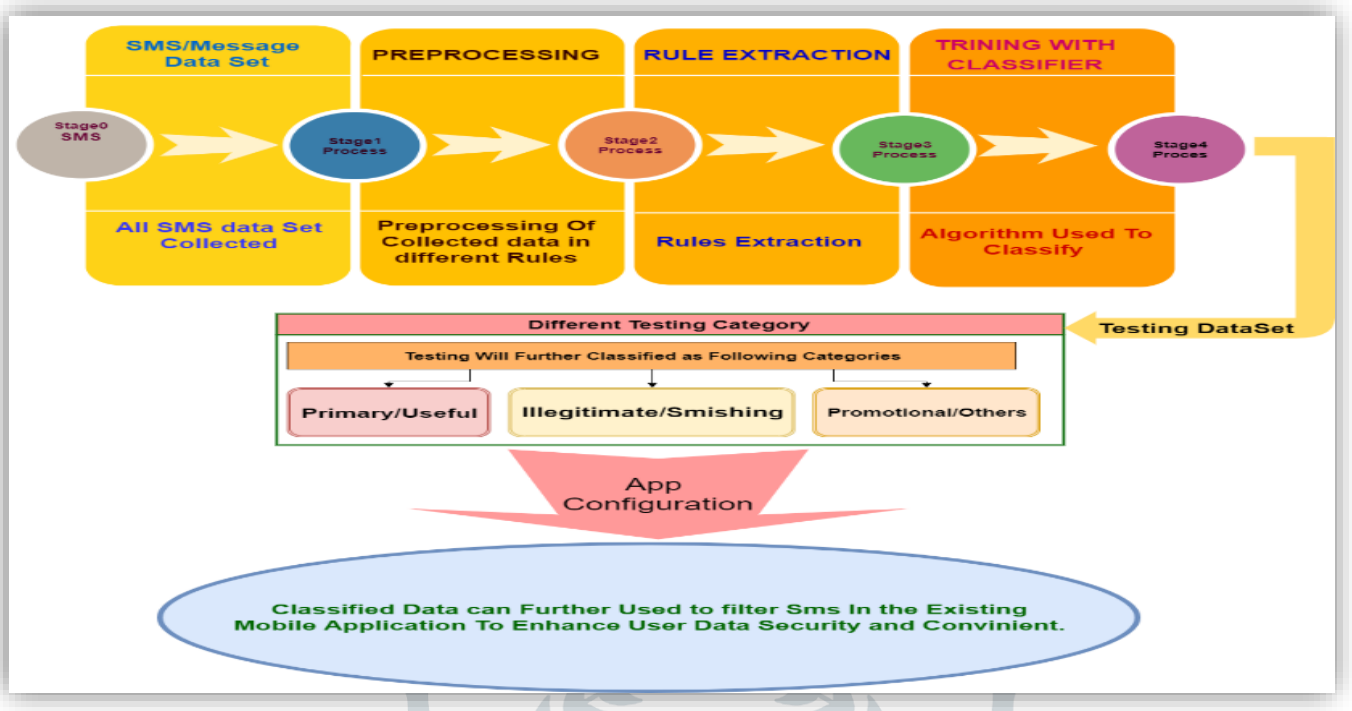
4 Tool and techniques

In our research work we used data mining classification techniques to classify the data and accordingly we found accuracy, and result set data, to do so we used WEKA tool, PC configuration having windows 10 core i5 processor 12 GB RAM etc.

Fig 3. Architecture of proposed

Steps will follow the different stages of data mining-

- Pre-processing
- Case- Extraction
- Classifier Training and Testing



Cases datasets to classify the datasets:

We have made more than 15 cases based on data collected from different sources, by using these cases we have analyses.

Fig4: Testing Data Classification Hierarchy

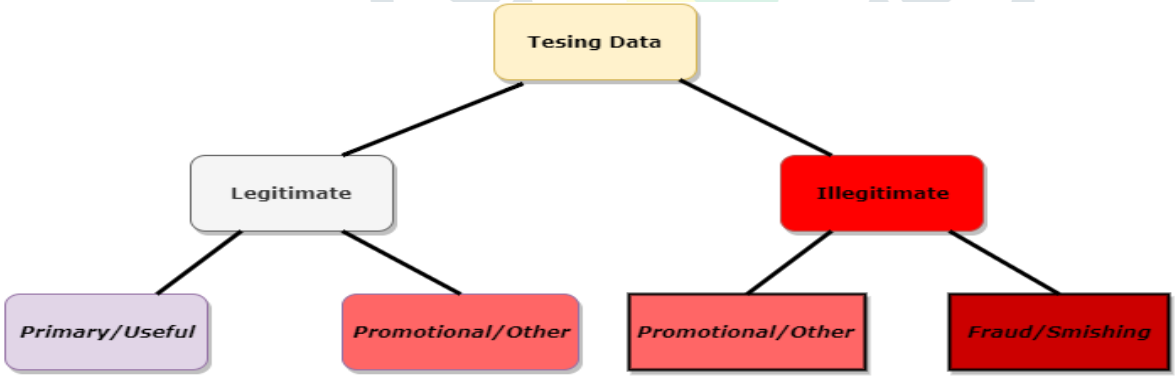
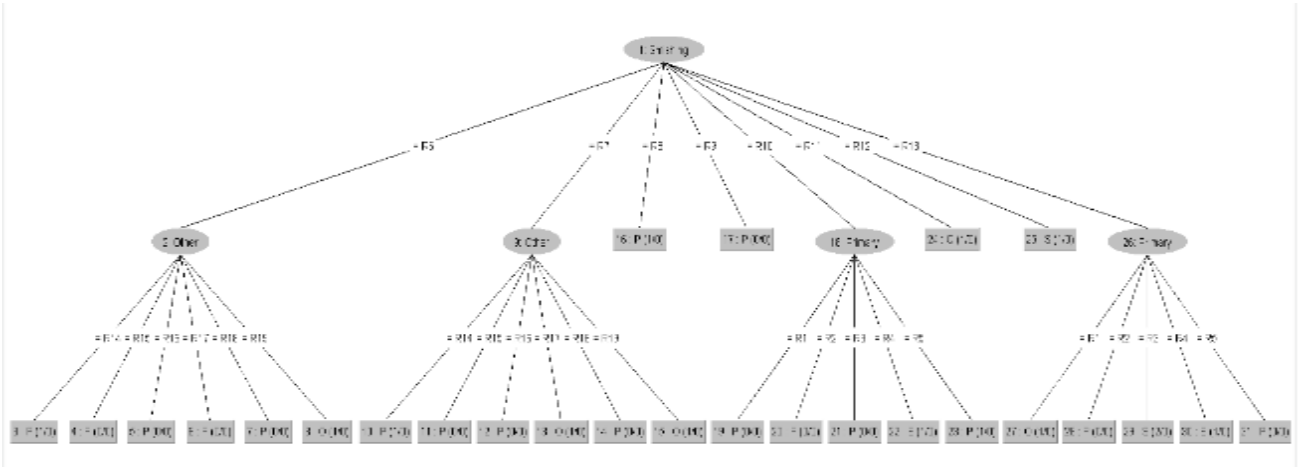


Fig 5: Different Algorithm analysis

Random Tree view



Random tree analysis

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     P
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     S
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     O
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

Bayes Net analysis

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     P
               1.000    0.111    0.833      1.000    0.909      0.861    1.000     1.000     S
               0.800    0.000    1.000      0.800    0.889      0.849    1.000     1.000     O
Weighted Avg.  0.929    0.040    0.940      0.929    0.928      0.896    1.000     1.000
```

Bayes Net Cost analysis

Confusion Matrix

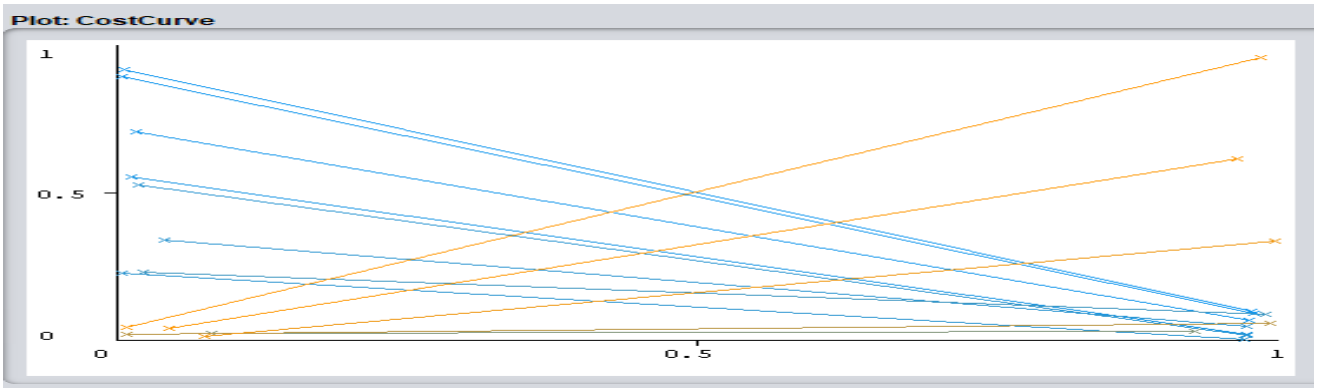
Predicted (a)	Predicted (b)	
0 (0%)	4 (28.57%)	Actual (a): P
0 (0%)	10 (71.43%)	Actual (b): S,O

Classification Accuracy: 71.4286%

Other measures:

Time taken to build model: 0.01 seconds
Time taken to Pre-process Data set: 0.01 seconds
Correctly Classified Instances 13 92.8571 %
Incorrectly Classified Instances 1 7.1429 %

Bayes Net Cost curve



5 Conclusion and future work

In our paper we have given an approach to secure SMS for now as well as in future. We had studied a lot with collected message data and made some cases to classify the messages into different category, so that it can be implemented into the existing application by the giant organizations like Google, Microsoft, and Apple not with different application but with the same existing application. Different inbuilt algorithms used by WEKA tool. True positive rate is 99.99%, and FPR is 0.01%.

As there is another application exist but giving the permission to the other application it's also a privacy breach so instead of securing system by other application it will be good to have a functionality in existing one application. Future research can have more data sets collected from different users, we will made more cases through it so that we can have a better security to the users. Also our planning to make more and more cases accordingly. We are exploring more algorithm and data set as well as cases and characteristics so that we can get better classification accuracy. Also we will extend this research work using data mining tools like Rapid Miner.

REFERENCES

[1] A. K. Jain and B. B. Gupta, A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security, 2016(9), 2016.

[2]Mobile User Statistics referred from here https://www.statista.com/statistics/274658/forecast-of-mobile-phone-users-in-india/

[3]SMS, C, The real value of sms to businesses, 2018, https://www. smscomparison.co.uk/sms-gateway-uk/2018-statistics/. (Accessed March 2019).

[4]T.A. Almeida, J.M.G. Hidalgo, A. Yamakami, Contributions to the study of sms spam filtering: new collection and results, in: Proceedings of the 11th ACM Symposium on Document Engineering, ACM, 2011, pp. 259–262.

[5]C. Wang, Y. Zhang, X. Chen, Z. Liu, L. Shi, G. Chen, F. Qiu, C. Ying, W. Lu, A behavior-based sms antispam system, IBM J. Res. Dev. 54 (2010) 3–1.

[6]T. Yamakami, Impact from mobile spam mail on mobile internet services, in: International Symposium on Parallel and Distributed Processing and Applications, Springer, 2003, pp. 179–184.

[7]V. Gupta, A. Mehta, A. Goel, U. Dixit, A.C. Pandey, Spam detection using ensemble learning, in: Harmony Search and Nature Inspired Optimization Algorithms, Springer, 2019, pp. 661–668.

[8]Z. Chen, Q. Yan, H. Han, S. Wang, L. Peng, L. Wang, B. Yang, Machine learning based mobile malware detection using highly imbalanced network traffic, Inform. Sci. 433 (2018) 346–364.

[9] C. Amrutkar, Y.S. Kim and P. Traynor, Detecting Mobile Malicious WebPages in Real Time, IEEE Transactions on Mobile Computing (2016)

[10] J.W. Joo, S.Y. Moon, S. Singh and J.H. Park, S-Detector: an enhanced security model for detecting Smishing attack for mobile computing, Telecommunication Systems vol. 66(1), 29–38 (2017).

[11] M. Foozy, C. Feresa, R. Ahmad and M.F. Abdollah, A practical Case based technique by splitting SMS phishing from SMS spam for better accuracy in mobile device, International Review on Computers and Software, vol. 9(10), pp. 1776-1782 (2014).

[12] E. M. El-Alfy and Ali A. AlHasan, Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm, Future Generation Computer Systems, vol. 64, pp. 98-107, (2016).

[13]Symantec Internet Security Threat Report, Available at: http://www.symantec.com/content/en/us/enterprise/other_resources/bistr_main_report_v19_21291018.en-us.pdf. Accessed August 2017

[14]Mobile messaging fraud report, Available at: https://mobileecosystemforum.com/mobile-messaging-fraud-report-2016/.

[15] Smishing Report, Available at : http://resources.infosecinstitute.com/category/enterprise/phishing/phishing-variations/phishing-variationssmishing/, last accessed 2017/07/15.

[16] The Social Engineering Framework, Available at: https://www.social-engineer.org/framework/attack-vectors/smishing/.

[17] J.W. Joo, S.Y. Moon, S. Singh and J.H. Park, S-Detector: an enhanced security model for detecting Smishing attack for mobile computing, Telecommunication Systems vol. 66(1), 29–38 (2017).

[18] M. Foozy, C. Feresa, R. Ahmad and M.F. Abdollah, A practical rule based technique by splitting SMS phishing from SMS spam for better accuracy in mobile device, International Review on Computers and Software, vol. 9(10), pp. 1776-1782 (2014).

[19] A. Tewari, A. K. Jain and B. B. Gupta, Recent survey of various defense mechanisms against phishing attacks. Journal of Information

[20] Dimensional Enterprise Mobile security Survey, Available at: http://blog.checkpoint.com/wpcontent/uploads/2017/04/Dimensional_Enterprise-Mobile-Security-Sury.pdf.

[21]N. Choudhary and A.K. Jain, Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique. Advanced Informatics for Computing Research, 18-30, 2017.

[22].A. K. Jain and B. B. Gupta, A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security, 2016(9), 2016.

[54]S.J. Delany, M. Buckley, D. Greene, Sms spam filtering: methods and data, Expert Syst. Appl. 39 (2012) 9899–9908.

[55]K. Mathew, B. Issac, Intelligent spam classification for mobile text message, in: Computer Science and Network Technology (ICCSNT), 2011 International Conference on, vol. 1, IEEE, 2011, pp. 101–105.

[60]I. Androulidakis, V. Vlachos, A. Papanikolaou, Fimess: filtering mobile external sms spam, in: Proceedings of the 6th Balkan Conference in Informatics, ACM, 2013, pp. 221–227.