SPAM MESSAGE IDENTIFICATION USING MACHINE LEARNING APPROACH

S.Mahammad Rafi Assistant Professor in Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India.

S.Venkata Padma Meghana, D.Sushma, P.Venkata Giridhar, T.Venkata Srinivasulu, A.C.Venkateshwara Reddy

Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India.

ABSTRACT:

We use some communication means to convey messages digitally. Digital tools allow two or more persons to coordinate with each other. This communication can be textual, visual, audio, and written. Smart devices including cell phones are the major sources of communication these days. Intensive communication through SMSs is causing spamming as well. Unwanted text messages define as junk information that we received in gadgets. Most of the companies promote their products or services by sending spam texts which are unwelcome. In general, most of the time spam emails more in numbers than Actual messages. In this paper, we have used text classification techniques to define SMS and spam filtering in a short view, which segregates the messages accordingly. In this paper, we apply some classification methods along with “machine learning algorithms” to identify how many SMS are spam or not. For that reason, we compared different classified methods on dataset collection on which work done by using the Weka tool.

Index Terms: Spam Messages, Classification, Spam Filtering, Comparison

I. INTRODUCTION:

In five years, there will be 3.8 billion mobile phone (smartphone) users, up from 1 billion . China, India, and the US are the top three countries in terms of mobile usage. Short Message Service, sometimes known as SMS, is a text messaging service that has been around for a while. You can use SMS services even without an internet connection. SMS service is thus accessible on both smartphones and low-end mobile devices. Although there are numerous text messaging apps on smart phones, such as WhatsApp, this service can only be used online. However, SMS is available at all times. Consequently, the need for SMS services is growing daily.

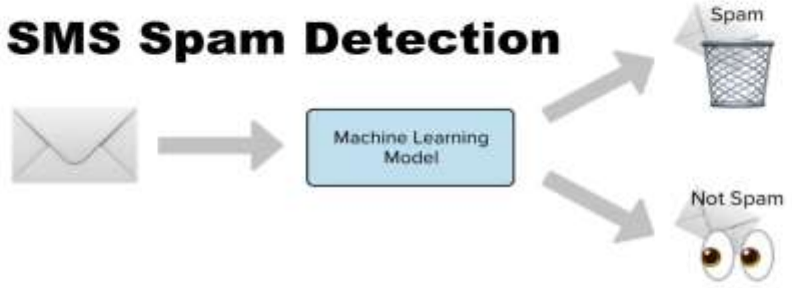


Figure 1: Spam Message Detection

Additionally, the system mostly derived knowledge from the data. For that goal, there are numerous methods available, including classification, clustering, and many others. SMS stands for short message service. 160-character messages must be sent by SMS, and lengthy messages must be broken up into numerous smaller messages. Short text messages could be exchanged between cell phones using the same communication protocols. The government intends to keep up with the rapid pace of technological progress. In previous years, the rate of text messaging climbed.

In some ways, SMS spam is sophisticated. The lowest SMS rates have made it possible for users and service providers to move away from the issue and limited availability of spam filtering apps for mobile devices. Email spam is less common than SMS spam. Even since it explains why 30% of typescript letters went to stylish Asia and 1% of transcript addressed in the United States. In 2004, the Telephone Customer Protection Act made SMS spam unlawful in the United States. Whoever receives unwelcome SMS knows how to lead the guidance counsellor to a court case with no real legal significance. Now in China, the top three mobile phone manufacturers agree to a joint plan to combat mobile spam by establishing limits on the number of typescript messages sent to one another since 2009.

In this paper, we demonstrate a few classification techniques that classify objects. Using classification techniques, we can determine whether a text message is spam or not. There must be a training set that contains the items at this establishment. SMS is text messaging. To summarise the SMS spam class or SMS into a human being, we employ text classification in this study. Since texts sent from people are considered to be from the human class or mobile phone, spam messages are typically provided by businesses and organisations to advertise their goods. Voice mail messages are frequently utilised as a communication tool since mobile phones and smartphones are generally used by a wide spectrum of people to make calls and send messages. Because SMS spam datasets are typically small in size, email filter spam contains more datasets than SMS spam when comparing the two types of spam. The filtering techniques scheme of the email spam filtering system could not be applied to SMS because of the small size of spam SMS. In certain nations, including Korea, email spam is less common than SMS spam. However, the opposite strategy was used in western regions, where email spam predominated over SMS spam because it was cheaper and more prevalent there. On mobile devices, about 50% of SMS messages are received as text messages, which are flagged as spam.

An SMS filtering system should function in reserve resources as well as in cell phone hardware because of this. We used ham and spam as real data in our analysis. We use a number of different categorization algorithms, some of which have been used in earlier research and some of which are brand-new.

Machine learning is a technology that allows computers to learn from the past and make predictions about the future. The majority of problems in the real world may now be solved using machine learning and deep learning in all fields, including health, security, market analysis, etc. Machine learning can be divided roughly into two types: Learning can be supervised, unsupervised, or semi-supervised.

One of the key subcategories of machine learning is supervised learning. Predictive modelling, another name for supervised learning, is the process of producing predictions from data. Classification and regression are two instances of supervised learning. supervised instruction For classification issues, the training data set has pre-labels, and for regression issues, function values are known. Switch to scoring later so that we can forecast values for fresh data when training is complete and the model has a minimum cost function for the training data set.

When a system receives a data set of emails, one supervised learning task is to determine if each email is spam or not. This is an example of supervised learning (ham). The fact that there is a predetermined result, such as spam or ham, makes this supervised learning. We applied multiple supervised learning techniques for SMS spam detection using a labelled dataset from UCI.

II. LITERATURE SURVEY:

Tiago, Almeida , José María GómezAkebo Yamakami. Contributions to the Study of SMS Spam Filtering. University of Campinas, Sao Paulo, Brazil.

The number of mobile phone users has increased, which has resulted in a sharp rise in SMS spam messages. The lower SMS usage rate, which has allowed many users and service providers to disregard the problem, as well as the restricted availability of mobile phone spam-filtering software make it challenging to combat mobile phone spam in practise. On the other hand, a significant disadvantage in academic contexts is the dearth of publicly available datasets for SMS spam, which are crucial for validating and contrasting various classifiers. Additionally, since SMS messages are typically brief, content-based spam filters may work worse.

We present the largest genuine, public, and unencrypted SMS spam collection we are aware of in this project. Additionally, we contrast the results of various tried-and-true machine learning techniques. The findings show that Support Vector Machine performs better than the other assessed classifiers, making it an excellent starting point for future comparison.

Duan, L., Li, N., & Huang, L. (2009). “A new spam short message classification” 2009 First International Workshop on Education Technology and Computer Science, 168-171.

This project suggests a method of message dual-filtering. The KNN classification algorithm and rough set are combined to first separate spam messages from other messages. It must re-filter some messages using the KNN classification algorithm to prevent lowering precision for reduction. Based on a basic set of the KNN classification algorithm, this method not only increases classification speed but also maintains excellent accuracy.

Inwhee Jo and Hyetaek Shim, "An SMS Spam Filtering System Using Support Vector Machine," Division of Computer Science and Engineering, Hanyang University, Seoul, 133-791 South Korea.

An SMS spam filtering system based on SVM (Support Vector Machine) and thesaurus is presented in this paper (Short Messaging Service). The system identifies words from sample data using a pre-processing tool, integrates their meanings using a thesaurus, derives features of integrated words using chi-square statistics, and then analyses these characteristics. The system has been tried out, and it works well in a Windows context

B. G. Becker. Visualizing Decision Table Classifiers. Pages 102- 105, IEEE (1998).

Decision trees, decision networks, and decision tables are all categorization models used for forecasting. Machine learning algorithms produce these. A decision table is made up of a hierarchy of tables where each entry is split down into its component parts by the values of two more characteristics to create a new table. Dimensional stacking is comparable to the structure [4]. Here, a visualisation technique is shown that enables even non-experts in machine learning to comprehend a model built on a variety of attributes. This representation is more practical than other static designs thanks to a variety of interactions.

Inwhee Joe and Hyetaek Shim, "An SMS Spam Filtering System Using Support Vector Machine," Division of Computer Science and Engineering, Hanyang University, Seoul, 133-791 South Korea:

This project describes a powerful and adaptive spam filtering system for SMS (Short Messaging Service) that uses SVM (Support Vector Machine) and a thesaurus. The system isolates words from sample data using a pre-processing device and integrates meanings of isolated words using a thesaurus, generates features of integrated words through chi-square statistics, and studies these features. The system is realized in a Windows environment and its performance is experimentally confirmed.

Spam filtering is a peculiar filed to automatic document classification to considering the document is spam or not. Automatic document classification means make bunch of similar documents by allocate each document to proper category by get through the classification system.. That classification is consisting of two phases.

First phase is feature selection method by extracting needed feature to classify after indexing bunch of documents. Second phase is decision make process that choose right category for the result from first phase. Automatic document classification gets ability to assign right category automatically through mechanical learning process.

For this process, it tagged specific word to bunch of learned document. The word represents the documents and extracting feature means batch job to select words revealed from learned document. However if it select every word in learned document as features, it takes too much time and looses judgment. To prevent this problem, calculate weight of information for each word then select featured words for automatic classification. In text categorization, we are dealing with a huge feature spaces. This is why; we need a feature selection mechanism. The most popular feature selection methods are document frequency thresholding (DF) , the X 2 statistics (CHI) , term strength (TS) , information gain (IG) , and mutual information.

Abhishek Patel#1 , Priya Jhariya*2 , SudalaguntaBharath#3 , Ankita wadhawan#4 “SMS Spam Detection using Machine Learning Approach”:

Spam is "unconstrained mass email" (Hidalgo, 2002), which "data made to be given to countless beneficiaries, notwithstanding their longings." Cormack (2007) depicted spam with propelling substance or compulsion content are passed on in the strategy for mass mailing Regardless, such spam could be unmistakable as demonstrated by the diverse media spam rehearses used, such email spam, SMS spam. Spammers flood the Short Message Service workers and give mass proportion of unconstrained Short Message Service to the end clients [16]. From a business point of view, Short Message Service clients need to contribute energy on destroying got spam Short Message Service which unquestionably prompts the advantage reduction and cause possible difficulty for affiliations. From this time forward, how to recognize the Short Message Service spam appropriately and proficiently with high precision changes into a gigantic report.

In this appraisal, information mining will be used to manage AI by utilizing various classifiers for preparing and testing and channels for information preprocessing and highlight choice. It plans to peer out the ideal mix model with higher precision or base on other metric's evaluation. As of now, there are various evaluation study done by utilizing information burrowing procedure for example, information digging by strategies for plan.

Altogether much exertion underscore on single classifier. In any case, spamming rehearses are changing the strategies to evade the spam territory [18]. Along these lines, in this examination, we will zero in on the whole around on framework for managing SMS spam by utilizing information mining technique. Questions, for example, regardless of whether the cross assortment model gives better precision result standing apart from any single classifier utilized for email spam unmistakable evidence will be seen through experimentation.

III. METHODOLOGY:

Using several supervised-learning techniques, the proposed method will focus on increasing the precision of spam message detection. Data was obtained via Kaggle. The dataset is then partitioned according to entropy. The fine-tuned dataset shows accuracy. Afterward, the divided dataset is used to observe accuracy. By using correlation and a working model, the optimal attributes for each leaf node are determined. The model's accuracy was seen to be hypertuned depending on the best attributes for each division. ML-based solutions have an advantage over blacklists because they can reduce the impact of zero-hour faked assaults, just like heuristic checks can. It's interesting to note that ML approaches can build their own categorization models by examining vast amounts of data. Due to ML algorithms' ability to discover their own models, manually creating heuristic tests is no longer necessary.

The analysis's module description is represented by the framework in Figure 2.

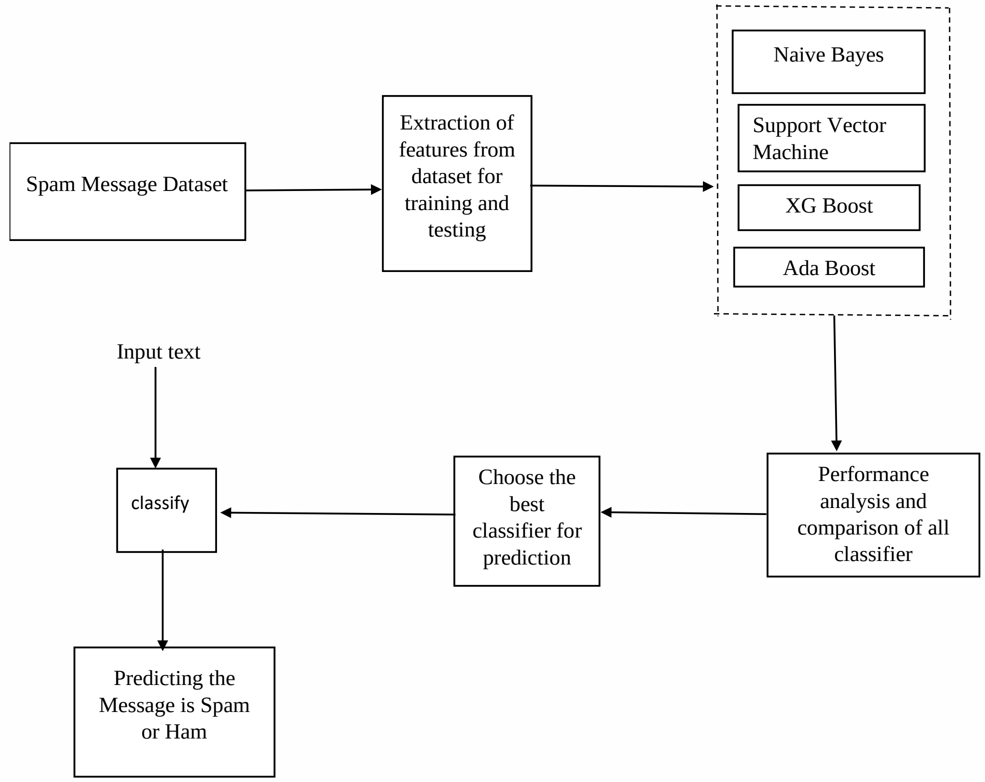

Spam Message Dataset
Extraction of features from dataset for training and testing
Naive Bayes
Support Vector Machine
XG Boost
Ada Boost
Input text
classify
Choose the best classifier for prediction
Performance analysis and comparison of all classifier
Predicting the Message is Spam or Ham

Fig 2: Block diagram for proposed system

A. Dataset

In this model, we've combined data sets that we've created with spam datasets that we've obtained from several online resources, including Kaggle. We test our model using 30% of the Kaggle spam dataset, and we train our model using the remaining 70%. Data from spam and genuine messages are included in the dataset.

B. Data preprocessing

The steps involved in data preprocessing include cleaning, instance selection, feature extraction, normalization, transformation, etc. The training dataset as a whole is the end outcome of data preprocessing. How data is pre-processed could have an impact on how the final results are understood. Filling in the gaps in the data, reducing noise, identifying and eliminating outliers, and resolving incompatibilities are all steps in the data cleaning process. The addition of certain databases or data sets may be accomplished through a technique called data integration. When collecting and normalizing data to measure a certain set of data, data transformation is taking place.

Data reduction allows for the creation of a very compact dataset overview that nevertheless contributes to the analysis's ability to yield a consistent result.

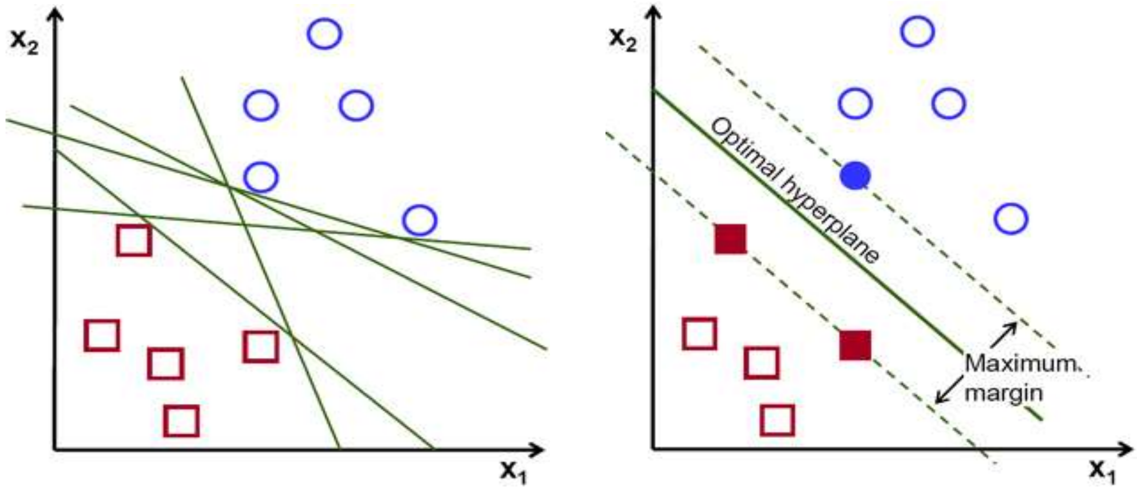
C. Train-test split

In order for the training dataset to be utilised to detect spam messages on the testing dataset, the dataset is divided into two subsets: testing set and training set. In order for the training model to adequately train and learn the data, 30% of the data is examined for the testing set.

IV. ALGORITHMS:

Support Vector Machine:

Support Vector Machine is a type of supervised machine learning algorithm that provides data analysis for classification and regression analysis. SVM is mostly used for classification. The value of each feature is equal to the value of the specified coordinate. Then, we detect the ideal hyperplane that differentiates between the two classes. Support vector machine is a representation as points in space contrasted into categories by a gap that is as wide as possible of the training data. It is effectual and efficient in high dimensional spaces and uses a subset of training points in the decision function, hence, it is also known for its memory efficiency. The algorithm indirectly provides probability estimations; these are calculated using five-fold cross-validation.



Naive Bayes :

A classification technique that is based on Bayes' Theorem with the presumption of independence among predictors. Naive Bayes is a way used to predict the class of the dataset. Using this, one can perform a multi-class prediction. If the assumption of independence is valid, then Naive Bayes is much more capable than the other algorithms like logistic regression. Furthermore, less training data is required for the classification. Naive Bayes classifier works efficiently in real-world situations such as document classification and spam filtering. Although, it is merely recognized as a bad estimator. It is an easy and a quick technique

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A/B) is the posterior probability of class (target) given predictor (attribute).

P(A) is the prior probability of class.

P(B/A) is the likelihood which is the probability of predictor given class.

P(B) is the prior probability of predictor.

XGBOOST:

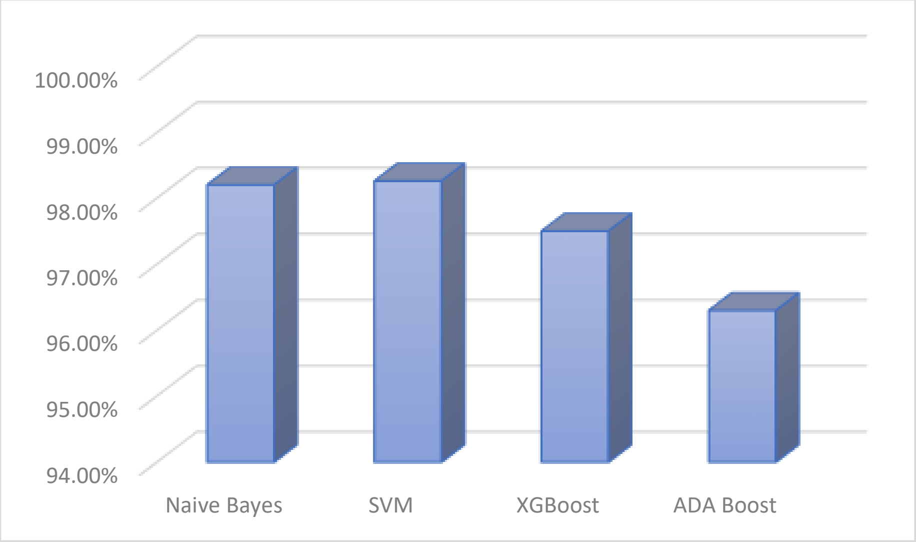
XG Boost stands for extreme Gradient Boosting. It is an application of gradient boosted decision trees, which is intended for its speed and performance. Boosting is an ensemble learning method where advanced techniques are included in order to rectify the errors made by the already proposed models. Models are included consecutively till we find that no additional enhancement can be carried out. While adding new models it uses a gradient descent technique to minimize the loss. The application of this algorithm is to provide efficient computational time and memory supplies. The aim of this design was to produce the best necessity of the accessible sources to train the model. Execution Speed and Model Performance are the two main reasons to work with XG Boost. This approach can support both classification and regression models.

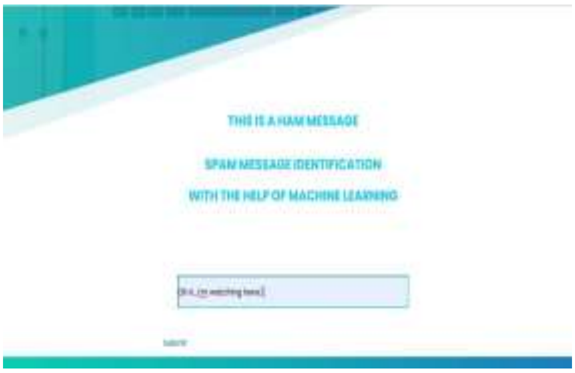
ADABOOST:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference.

V. EXPERIMENTATION AND RESULTS:

In this paper we applied Hashing Vectorizer word embedding technique and then applied four classification algorithms. The results of the experiments are shown in table. We achieved best accuracy with SVM . After Applying the algorithms it is used to identify whether the Message is Ham or Spam.



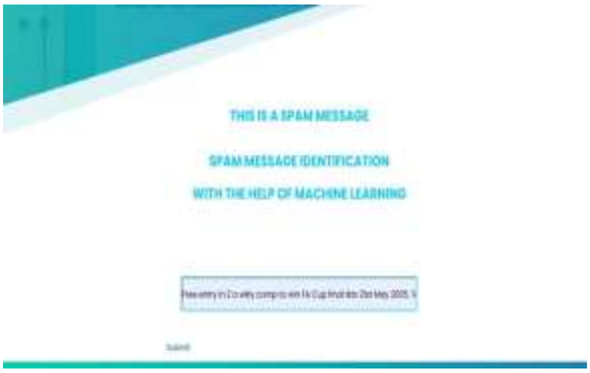


THIS IS A HAM MESSAGE

SPAM MESSAGE IDENTIFICATION

WITH THE HELP OF MACHINE LEARNING

Oh k...i'm watching here:)



THIS IS A SPAM MESSAGE

SPAM MESSAGE IDENTIFICATION

WITH THE HELP OF MACHINE LEARNING

Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. T

VI. CONCLUSION:

We presented a spam message categorization utilising several algorithms such as Naive Bayes ,Support vector machine, XGBoost and ADABoost . For the assessment of spam base datasets using the Weka tool, two classification techniques are employed in Weka: cross validation and training set. The same data will be utilised for training and testing in the training set. In addition, for cross validation, training data is separated into many folds. Following implementation and experimental analysis, get the result that classifier with training set provides accuracy. As a consequence, Support vector machine is the strategy that produces the best results for spam msg categorization. On just one dataset, we tested this model. Future tests of our model on various datasets are planned.

VII. REFERENCES

[1] J. Han, M. Kamber. Data Mining Concepts and Techniques. by Elsevier inc., Ed: 2nd, 2006

[2] A. Tiago, Almeida , José María GómezAkebo Yamakami. Contributions to the Study of SMS Spam Filtering. University of Campinas, Sao Paulo, Brazil.

[3] M. Bilal Junaid, Muddassar Farooq. Using Evolutionary Learning Classifiers To Do Mobile Spam (SMS) Filtering. National University of Computer & Emerging Sciences (NUCES) Islamabad, Pakistan.

[4] Inwhee Joe and Hyetaek Shim, "An SMS Spam Filtering System Using Support Vector Machine," Division of Computer Science and Engineering, Hanyang University, Seoul, 133-791 South Korea.

[5] Xu, Qian, Evan Wei Xiang, Qiang Yang, Jiachun Du, and Jieping Zhong. "Sms spam detection using noncontent features." IEEE Intelligent Systems 27, no. 6 (2012): 44-51.

[6] Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., and Naik, V. "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering," Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, ACM, 2011, pp. 1-6.

[7] Duan, L., Li, N., & Huang, L. (2009). "A new spam short message classification" 2009 First International Workshop on Education Technology and Computer Science, 168-171.

[8] Weka The University of Waikato, Weka 3: Data Mining Software in Java, viewed on 2011 September 14.

[9] Mccallum, A., & Nigam, K. (1998). "A comparison of event models for naive Bayes text classification". AAAI-98 Workshop on 'Learning for Text Categorization'

[10] Bayesian Network Classifiers in Weka, viewed on 2011 September 14.

[11] Llora, Xavier, and Josep M. Garrell (2001) Evolution of decision trees, edn., Forth Catalan Conference on Artificial Intelligence (CCIA2001).

[12] B. G. Becker. Visualizing Decision Table Classifiers. Pages 102- 105, IEEE (1998). A. Bantukul and P. J. Marsico, ''Methods, systems, and computer program products for short message service (SMS) spam filtering using E-mail spam filtering resources,'' U.S. Patent 7 751 836 B2, Jul. 6, 2010.

[13] H.-Y. Chou and N.-H. Lien, ''Effects of SMS teaser ads on product curiosity,'' Int. J. Mobile Commun., vol. 12, no. 4, pp. 328–345, Jul. 2014.

[14] N. Jindal and B. Liu, ''Review spam detection,'' in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 1189–1190.

[15] M. Jiang, P. Cui, and C. Faloutsos, ''Suspicious behavior detection: Current trends and future directions,'' IEEE Intell. Syst., vol. 31, no. 1, pp. 31–39, Jan./Feb. 2016.

[16] C. Wang et al., ‘‘A behavior-based SMS antispam system,’’ IBM J. Res. Develop., vol. 54, no. 6, pp. 3:1–3:16, Nov./Dec. 2010.