POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)

URL: http://dergipark.org.tr/politeknik

Support vector machine based spam SMS detection

Destek vektör makineleri kullanılarak spam SMS tespiti

***Yazar(lar) (Author(s))**: Adem TEKEREK*

ORCID: 0000-0002-0880-7955

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Tekerek A., "Support vector machine based spam sms detection", ***Politeknik Dergisi***, 22(3): 779-784, (2019).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.429707

Support Vector Machine Based Spam SMS Detection

Araştırma Makalesi / Research Article

Adem TEKEREK[*]

Information Technology Department, Gazi University, Turkey

(Geliş/Received : 02.10.2016 ; Kabul/Accepted : 26.08.2017)

ABSTRACT

Short Message Service (SMS) is the most important communication tool in recent decades. With the increased popularity of mobile devices, the usage rate of SMS will increase more and more in years. SMS is a practical method used to reach individuals directly. But this practical and easy method can cause SMS to be misused. The advertising or promotional SMS of the companies are an examples of this misuse. In this study, a spam SMS detection technique is proposed using Data Mining (DM) methods. In the proposed study, data mining algorithms such as Naive Bayes (NB), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Random Forest (RF) and Random Tree (RT) is selected. SMSSpamCollection dataset, which is contain 747 spam SMS and 4827 ham SMS, is used. 10 fold cross-validation technique is used to evaluate prediction of Spam SMS in the dataset. Therefore, proposed study achieved 98.33 % success rate and 0,087 false positive rate for SVM algorithm..

Keywords: Spam SMS, data mining, machine learning, support vector machine.

Destek Vektör Makineleri Kullanılarak Spam SMS Tespiti

ÖZ

Kısa Mesaj Servisi (SMS) son yılların en önemli iletişim araçlarından biri haline gelmiştir. Mobil cihazların artan popülaritesiyle, SMS kullanım oranları da yıllar içinde daha da artmaya devam edecektir. SMS doğrudan bireylere ulaşmak için kullanılan pratik bir yöntem olarak kullanılmaktadır. Ancak bu pratik ve kolay yöntem, SMS'in yanlış ve kötü amaçlı kullanılmasına da neden olabilmektedir. Şirketlerin reklam veya tanıtım SMS'leri bu yanlış kullanımın önemli örneklerindendir. Bu çalışmada, Veri Madenciliği (DM) yöntemleri kullanılarak bir spam SMS tespit tekniği önerilmiştir. Önerilen çalışmada, Naive Bayes (NB), K-En Yakın Komşuluk (KNN), Destek Vektör Makinesi (SVM), Rastgele Orman (RF) ve Rastgele Ağaç (RT) gibi veri madenciliği algoritmaları seçilmiştir. Bu çalışmada 747 spam SMS ve 4827 jambon SMS içeren SMSSpamCollection veri kümesi kullanılmıştır. Veri kümesindeki Spam SMS tahminini değerlendirmek için 10 katlı çapraz doğrulama tekniği kullanılmıştır. Önerilen yaklaşımda Destek Vektör Makineleri sınıflandırma algoritması ile %98.33 oranında başarılı tespit yapılarak, 0,087 yanlış pozitif oran elde etmiştir.

Anahtar Kelimeler: İstenmeyen SMS veri madenciliği, makine öğrenmesi, destek vektör makineleri

1. INTRODUCTION

Short Message Service (SMS) is one of the most popular communication service where messages are sent electronically. The increase in the use of mobile devices also increased the number of SMS sent and received. With the increased use of SMS, the cost of SMS services has also decreased. The low price and the high bandwidth of the SMS network have attracted a large amount of SMS spam. This rise has also increased the malicious use of SMS, resulting in a spam SMS problem. A spam SMS is any unwanted message that is sent to user's mobile devices. Spam SMS include advertisements, free services, promotions, etc. According to people classify SMS Spam as annoying (32.3%), wasting time (24.8%), and violating personal privacy (21.3%) [1].

SMS is not text-rich. Therefore, spam SMS detection is generally based on text mining. Text mining aims to get structured data through the text, such as classification,clustering, concept or entity extraction, texts production of granular taxonomy, textual sentimental analysis, document summarization and entity relationship modeling. To obtain the structured data, information retrieval, lexical analysis, pattern recognition, word frequency, tagging, information extraction, data mining and visualization methods are used [2].

Several approaches such as machine learning and data mining have been used in spam SMS detection.

El-Alfy et all. [3] proposed a model for filtering messages for both SMS and email. They analyzed different methods in order to finalize features set such that complexity can be reduced. Authors have used SVM and Naive Bayes algorithms, and features which are URLs, spam words, emotion symbols, spam domain, special characters, defect words, metadata, Javascript, function words, recipient address and subject field. They have evaluated their proposed model on five email and SMS datasets.

Chan et al. [4] proposed two methods for SMS spam detection, feature rewriting and good word attack. These methods focus on the length of the SMS along with the length of the SMS. A good word attack provides a new

[*] *Sorumlu Yazar (Corresponding Author)*
e-posta : atekerek@gazi.edu.tr

rescaling function to re-scale weights while focusing on classifying the classifier using the least number of characters in the rewrite method. The authors evaluated the experiment on datasets, SMS and comments.

Delany et al. [5] discuss different approaches available for SMS Spam filtering and the problems associated with the dataset collection. They analyzed a large dataset of SMS spam and used ten clusters such as ringtones, prizes, services, finance, competitions, dating, claims chat, voicemail and others.

Xu et al. [6] have identify SMS Spam messages using content features. They used SVM and k-nearest neighbor and static, temporal and network features. They found that by combining temporal and network features SMS Spam messages can be detected more accurately and with good performance. They found the filter SMS Spam messages by using features that contain temporal information and graph-topology thus excluding the content of the message.

Yadav et al. [7] proposed a SMSAssassin model for SMS filtering. They used a feature set of 20 lightweight features and two machine learning algorithms, SVM and Bayesian learning. They used a collected a dataset of 2000 messages. In authors proposed model whenever the user gets some SMS over his phone, then SMSAssassin captures that SMS without user's knowledge, fetches feature values, and sends these values to the server for classification. If the messages are marked as spam SMS, the user will not see this message and it will be directed to the spam folder.

Hidalgo et al. [8] have analyzed that how Bayesian filtering technique can be used to detect SMS Spam. They have collected two datasets one in English and another in Spanish. Their analysis shows that Bayesian filtering techniques that were earlier used in detecting email spam can also be used to block spam SMS.

Almeida et al. [16] proposed a comparison study by using some supervised learning algorithms to give results for each one. They used a dataset which contains 747 spam SMS and 4,827 non-spam SMS. They used 13 classifiers for experiment include 8 variations of SVM, Naive Bayes, Description Length classifier, k-NN, C4.5, a rule learner. According the experimental result, linear SVM performs best, with an overall accuracy of 97,64%, a false positive rate of 0,18%.

In this study, feature selection is implemented automatically from the dataset. Thus, the detection of spam SMS messages is more precise. Data mining algorithms such as Naive Bayes (NB), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Random Forest (RF) and Random Tree (RT) is used for best detection of spam SMS. The selection of SVM as a classification technique has increased the success of detection spam SMS.

This paper consists of five sections. In section 2 the data mining techniques, used in this study are explained. In section 4, dataset and feature extraction is detailed. In section 5 proposed spam detection model is presented. In section 6, conclusion is given.

2. MATERIAL and METHOD

In this study, some data mining classification methods have been used to determine if a SMS is actually a spam SMS or ham SMS. Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF) and Random Tree (RT) data mining methods are used to classify SMS as malicious or not. These methods are described below.

2.1. Naive Bayes (NB)

Naive Bayes is one of the most effective learning algorithms in machine learning. Bayesian spam SMS filtering is a statistical method of detecting spam SMSs based on Bayes' theorem to calculate the probability that a SMS is actually a spam SMS. Naive Bayes algorithm is one of the machine learning methods that is used in text classification. It is a statistical inference based on probability, and is used to determine previously created classes [9].

NB uses a discriminant function to compute the conditional probabilities of P(Ci|X). As shown in formula (1) the inputs, P(Ci | X) denotes the probability that, example X belongs to class Ci

$$P(C_i|X) = \frac{P(C_i)*P(X|C_i)}{P(X)} \qquad (1)$$

P(Ci) is the probability of observing class i. P(X | Ci) denotes the probability of observing the example, given class Ci. P(X) is the probability of the input, which is independent of the classes.

2.2. K-Nearest Neighborhood (KNN)

The KNN is a pattern classifier that allows classification without the need to know the probability distributions of classes [10]. In this method, the distance of each test vector from the set of training vectors to be classified is calculated. At the next stage of the test vector, the class is assigned to the k majority of the closest vectors. The success of this method affects the selected distance measure and usually the experimentally determined k value.

2.3. Support Vector Machine (SVM)

In this study, SVM classification method is used to determine if a SMS is a spam SMS or ham SMS. SVM data mining method is used to classify SMS as malicious or not. SVM is a linear separation limit (wTx + b = 0) that classify the samples correctly. SVM, a supervised learning technique, is a combination of a linear machine learning technique. SVM is a two-dimensional variable class forms a hyperplane that divides the margin between hyperplane and the nearest data points by maximizing the weight vector w to the feature vector [11]. SVM decision boundary scheme is presented in Figure 1.

The most important advantage of the SVMs is the classification problem is squared optimization problem. Therefore, to solve the problem the number of transactions decreases during the learning phase and other techniques algorithms [12].

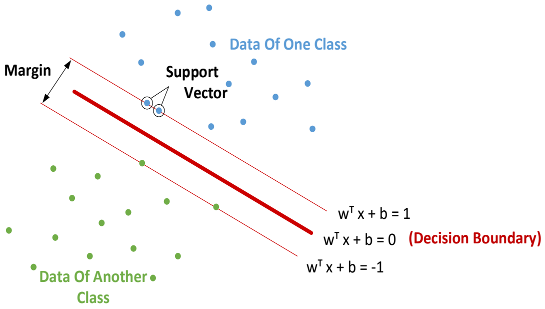


Figure 1. SVM decision boundary scheme

SVM is based on the concept of decision plans that define decision boundaries. A decision plane is a decision between a set of objects with different class memberships and the SVM modeling algorithm finds the best hyperplane with the maximum margin to separate from the two classes, which requires the following optimization problem to be solved.

Maximum ;

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(X_i, X_j) \qquad (2)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (3)$$

where $0 \leq \alpha_i \leq b$, i = 1,2,..,n

In equation (1) and equation (2) α_i is the weight of training sample x1. If $\alpha_i > 0$, x_1 is called a support vector b is a regulation parameter used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be achieved. K is a kernel function, which is used to measure the similarity between two or much more samples [13].

2.4. Random Forest (RF)

RF is a controlled machine learning algorithm that creates a forest and makes it random. RF is a type of decision tree algorithm, and trained by bagging method. The purpose of the bagging method is to increase the result of a combination of learning models. RF can be used for both classification and regression problems. Classification can be considered the building block of machine learning. RF adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. In Figure 2, scheme of random forest is given.

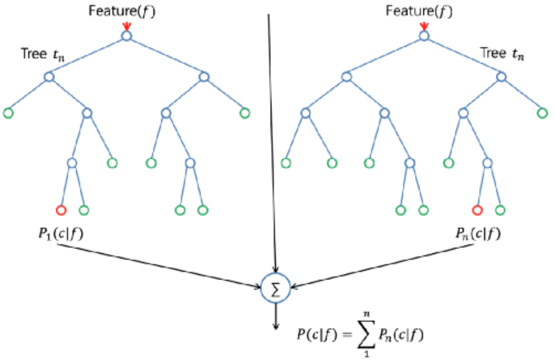


Figure 2. Random Forest scheme

Therefore, in RF, only a random subset of the features is taken into consideration by the algorithm for splitting a node. It can be even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds [14].

2.5. Random Tree (RT)

RT is a tree or arborescence that is formed by a stochastic process. RT is generation of a variety of trees at "random," and for small numbers of leaves it can generate all possible trees [15]. Random trees have several usage areas; used for phylogeny programs that do not have the ability to examine all trees or clusters of random trees;

- Used for estimate distributions of tree comparison measures,
- Used for the production of all possible tree shapes,
- Used as a basis for statistical tests.

In Figure 3, RT working scheme is presented.

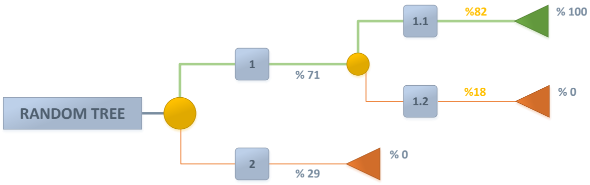


Figure 3. RT working scheme.

3. DATASET AND FEATURE EXTRACTION

In this study, a dataset consisting of 4.827 legitimate SMS and 747 spam SMS produced by T. A. Almeida is used. This is the largest existing SMS spam dataset currently available. Table 1 shows the basic statistics of the dataset. SMS spam collection dataset is a large real, public and non-encoded SMS dataset spam collection which was proposed by Almeida et. al. [16].

Table 1: Basic statistics

Messages	Amount	
	Messages	Percent (%)
Hams	4,827	86,60
Spams	747	13,40
Total	5,574	100,00

Table 2 presents the statistics related to the tokens extracted from the corpus. Note that, the proposed dataset has a total of 81,175 tokens and mobile phone spam has in average ten tokens more than legitimate messages.

Table 2: Token statistics

Token statistics hams	63,632
Spams	17,543
Total	81,175
Avarage per message	14.56
Avarage in hams	13.18
Avarage in spams	23.48

4. PROPOSED APPLICATION

Proposed study, main aim is to filter the spam and ham SMS using data mining algorithms. In Figure 4, the steps of the proposed study is given. At first, dataset is selected and pre-processed. After pre-processing, features are automatically generated using the dataset. So classification was performed using different data mining algorithms. According to the evaluation results SVM gave the best result.

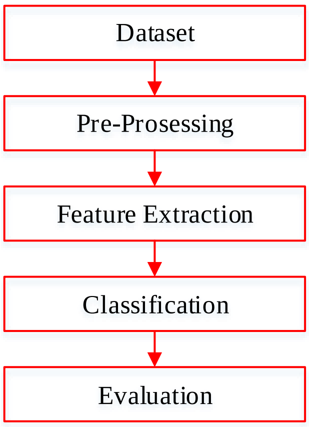


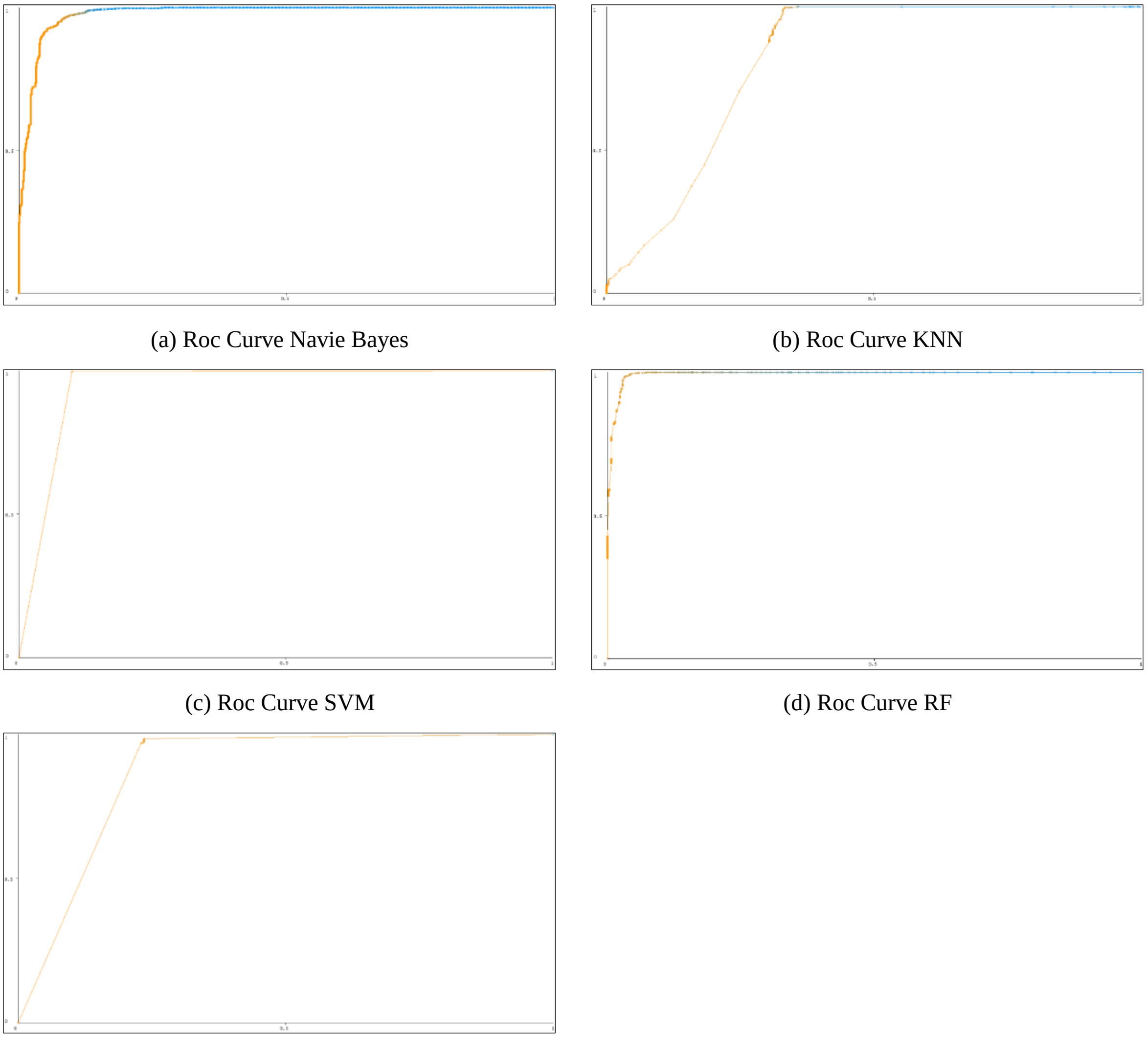
Figure 4. Proposed study working scheme.

In this study text mining based detection is proposed. The detection rules are parsed by text (token) mapping of the SMS text content. Specification of the rules required for the detection of spam SMS has been established. Each rule performs a test on the SMS dataset, and each rule has a score for decision about whether a SMS is spam or not. If the results exceed the threshold, then the SMS is marked as spam and the others are classified as ham SMS. The results obtained in the study are compared with the studies [1,16] as seen in Table 3. According to the results of the comparison, proposed model has the best performance with SVM. SVM has the highest True Positive (TP) (0,983) and the lowest False Positive (FP) (0,087) rate.

Table 3. Evolutionary Results

Order	Study	Dataset	Method	Result(%)	FP	TP
1	Proposed Study	SMS Spam Collection	**SVM**	**98.3315**	**0,087**	**0,983**
			NB	96.7887	0,108	0,968
			RT	95.4611	0,206	0,955
			RF	97.4345	0,166	0,974
			KNN	95.1381	0,310	0,951
2	Choudhary, N., et al [1]	SMS Spam Collection	NB	94.1	0.077	0.941
			DT	96	0.133	0.960
			RF	96,5	0.102	0.965
3	Almeida, T.A., el al [16]	SMS Spam Collection	linear SVM	97,6	0.18	0.976

ROC Curve graph of data mining methods used in the study is presented in Figure 7. In Figure 7 ham SMS represents zero (0), spam e-mails represents one (1). Figure 7 (c) represents the ROC Curve graph of the SVM classification method that yields the most successful results.



(a) Roc Curve Navie Bayes

(b) Roc Curve KNN

(c) Roc Curve SVM

(d) Roc Curve RF

(e) Roc Curve RT.

Figure 7. Roc curve graphs of classification methods

5. CONCLUSION

In this study, it is aimed to detect spam SMS using some data mining algorithms, which are Naive Bayes (NB), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Random Forest (RF) and Random Tree (RT). Although there are many SMS spam filtering studies, due to the existence of spammers and adoption of new techniques, SMS spam filtering becomes a challenging problem to the researchers. The dataset which was developed by Almeida et. al. [16]. The performance of proposed model was evaluated using training set and observed that SVM classifier outperforms than other classifiers and the false positive rate is also very low compared to other studies and other algorithms used in this study. Also 10 fold cross-validation technique is used to evaluate. SMS spam filters using this approach can be adopted either at SMS server or at SMS client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth, and storage usage.

REFERENCES

[1] Choudhary, N., & Jain, A. K., “Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique” ***In Advanced Informatics for Computing Research Springer, Singapore,*** 18-30 (2017).

[2] Mujtaba, G., & Yasin, M., “SMS spam detection using simple message content features”, ***J. Basic Appl. Sci. Res,*** 4(4),:275-279, (2014).

[3] El-Alfy, E.S.M., AlHasan, A.A., “Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm”, ***Future Gen. Comput. Syst***. 64: 98–107, (2016).

[4] Chan, P.P.K., Yang, C., Yeung, D.S., Ng, W.W.Y., “Spam filtering for short messages in adversarial environment”, ***Neurocomputing***, 155: 167–176, (2015).

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

[6] Xu, Q., Xiang, E.W., Yang, Q., Du, J., Zhong, J., “SMS spam detection using non-content features”, ***IEEE Intelligent Systems,*** 27(6): 44–51, (2012).

[7] Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., Naik, V., “SMSAssassin: crowdsourcing driven mobile-based

system for SMS spam filtering", ***12th Workshop on Mobile Computing Systems and Applications,*** 1–6, (2011).

[8] Hidalgo, J.M.G., Bringas, G.C., Sánz, E.P., García, F.C., "Content based SMS spam filtering", ***In ACM Symposium on Document Engineering***, 107–114, (2006).

[9] Awad, W. A., & ELseuofi, S. M., "Machine Learning methods for E-mail Classification", ***International Journal of Computer Applications***, 16(1): (2011).

[10] S. Theodoridis and K. Koutroumbas, Pattern Recognition, ***Fourth Edition: Academic Press***, (2008).

[11] G. Chechik, G. Heitz "Max-margin Classification of Data with Absent Futures", ***In Journal of Machine Learning Research,*** 9: (2008).

[12] Osowski, S., Siwekand, K., and Markiewicz, T. "MLP and SVM Networks – a Comparative Study", ***Proceedings of the 6th Nordic Signal Processing Symposium***, (2004).

[13] Arana-Daniel, N., Gallegos, A. A., López-Franco, C., Alanís, A. Y., Morales, J., & López-Franco, A., "Support vector machines trained with evolutionary algorithms employing kernel adatron for large scale classification of protein structures", ***Evolutionary Bioinformatics***, 12: (2016).

[14] Strobl, C., Malley, J., & Tutz, G., "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests", ***Psychological methods***, 14(4): 323, (2009).

[15] Lladó, A., "Decomposing almost complete graphs by random trees" ***Journal of Combinatorial Theory,*** **Series A**,154: 406-421, (2018).

[16] Almeida, T.A., Gomez Hidalgo, J.M., Yamakami, A., "Contributions to the Study of SMS Spam Filtering: New Collection and Results", ***Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11),*** Mountain View, CA, USA, (2011).