Jestr

Journal of Engineering Science and Technology Review 10 (1) (2017) 131- 137

Research Article

JOURNAL OF
**Engineering Science and Technology Review**

www.jestr.org

# SMS Spam Filtering Using N-gram method, Information Gain Metric and an Improved Version of SVDD Classifier

**Mohamed El Boujnouni**

*Mohammed V University, Faculty of Sciences Rabat, Laboratory of Conception and Systems (Microelectronic and Informatics) Avenue Ibn Battouta B.P 1014, Rabat, Morocco*

Received 2 January 2017; Accepted 17 March 2017

***Abstract***

Text messages (SMS or Short Message Service) are widely used form of mobile communication; their popularity is attributed to many factors including low cost sending, simple delivery mode and convenient usage. However, an unsolicited kind of SMS (Spam) has been appeared and caused major problems for users and mobile service providers. In this paper we propose a new SMS Spam filter able to distinguish between legitimate messages and Spam. The proposed filter is based on three components: N-grams method to extract features from short messages, information gain ratio to select the most relevant features, and an improved version of Support Vector Domain Description to detect SMS Spam. Experimental results on a large benchmark dataset of real-world benign and Spam SMS have shown the performance and effectiveness of the proposed filter.

*Keywords:* Short Message Service, N-gram Method, information gain criterion, Support Vector Domain Description classifier

## 1. Introduction

"Congrats! 1 year special cinema pass for 2 is yours". Don't believe on this text, it's just a part of a Spam message taken from a public SMS Spam collection [4]. It is undoubtedly that receiving a massive quantity of SMSs of such type at inappropriate times of day, will be very annoying for mobile users especially when they contain very attractive fake offers. Short Message Service known by its acronym SMS is a mechanism to deliver short messages over mobile networks. As its name indicates, SMS can deliver very limited data namely one SMS can contain at most 140 bytes (1120 bits) of data which is limited to either 70 or 160 characters in length. SMS has many technical attractive proprieties including: It can handle with all languages supported by Unicode, also it can carry binary data in addition to text (i.e pictures, logo), and furthermore SMS is supported by all Global System for Mobile (GSM) phones.
Like email, SMS is prone to Spamming. The latter can be defined as the abuse of phone text messaging service, Web, or mobile communication systems to send unwanted messages in large quantities often for commercial goals to an indiscriminate set of recipients. SMS Spam can contain a simple text message, link(s) to a phone number to call, link(s) to a website, binary data, etc. SMS Spam can cause various damages including, but not limited to, spending the storage space in Sh ort Message Service Center (SMSC) since SMS is a store-and-forward service, wasting the mobile network bandwidth by receiving massive quantity of Spam, filling the user's phone inbox by saving unsolicited SMS, consuming user's time and efforts by distinguishing between Spam and benign SMS, hunting for sensitive information by imitating official SMS delivered by trusted authorities, leading to financial losses to users and mobile service providers, breaking the law by delivering prohibited contains. Spam problem can be overcome with various techniques such as: Applying the law against Spammers, although it's difficult to track the actual offenders, making a blacklist (vs whitelist) of contacts. Senders existing in this list are considered as Spammers, and their messages are blocked while messages from senders in a whitelist are considered as legitimates, and thus their texts are delivered whatever their content, using digital signatures to distinguish between legitimate SMS and Spam. The signatures can be provided by SMS receivers or mobile service providers, filtering SMS basing on users' behaviors. When several users identify a message as unsolicited the service provider considers that message as Spam, analyzing the content of suspect SMSs by searching for special signs used by spammers. Those signs include indicative words, unusual distribution of punctuation marks, numbers, capital letters, etc.

In this paper we propose a new approach to filter automatically unsolicited SMS messages basing on their content. The proposed technique is based on three components: Character-level N-grams to extract features from SMS, information gain (IG) to select the most relevant features and an improved version of Support Vector Domain Description (SVDD) to distinguish Spam from Ham messages. The idea is motivated by the fact that N-grams and IG are two excellent techniques widely used for text categorization, and SVDD is a powerful classifier that has shown good performance in many challenging classification problems.

E-mail address: med_elbouj@yahoo.fr

This paper is organized in the following way: Section 2 presents an initial background of content-based SMS Spam filtering. Section 3 presents in details different components of the proposed filter including: N-gram method, Information gain metric and the improved version of Support Vector Domain Description. Section 4 begins by introducing a benchmark dataset widely used in SMS Spam detection researches to evaluate the accuracy of filters. Then, it presents a series of experiments designed to test the performance of the proposed filter on this dataset and finally section 5 provides main conclusions.

## 2. Related works

In recent years, several Spam filters based on machine learning and data mining techniques have been proposed and tested on various databases of SMS. Hidalgo et al. [1] evaluated the capability of a number of classification algorithms and text representation methods to detect SMS Spam. After evaluating experimentally their work using two different SMS datasets of Spanish and English messages, they concluded that Bayesian filtering technique can be employed successfully to detect SMS Spam. Healy et al. [2] compared the performance of detecting SMS Spam using three classifiers K-Nearest-Neighbor (KNN), Support Vector Machines (SVM), and Naive Bayes. They concluded that SVM and Naïve Bayes significantly outperform the KNN classifier. Nuruzzaman et al. [3] studied the possibility of detecting SMS Spam on mobile phones taking into consideration their technical limitations in terms of storage, memory and CPU. Their filter, based on text categorization techniques, was able to detect SMS Spam with reasonable accuracy, minimum storage consumption, and acceptable processing time. Longzhen et al. [5] presented a multi-filtering approach to detect SMS Spam. Their approach operates in two steps: Initially rough sets method was applied to filter a given short message. If the message was classified as Spam then it's passed to k-NN classifier for final confirmation. The authors found that their approach not only improves the speed of classification but also retains high accuracy. Mahmoud et al. [7] proposed an anti-spam filter based on Artificial Immune System (AIS). The aforesaid filter includes: tokenizer, analysis engine, stop word filter, AIS engine, and training process. The experimental results have shown that their filter can classify SMS Spam and ham with accurate compared to Naïve Bayesian algorithm. Joe et al. [13] presented an SMS Spam filtering system based on SVM classifier and a thesaurus. The system works in four steps: At first it extracts words from messages then it searches the meanings of these words using a thesaurus, next it generates features from these words through chi-square statistics and finally it classify the features using SVM classifier. The performance of their system was experimentally evaluated. Uysal et al. [14] designed a novel framework for SMS spam detection. The framework finds out discriminative features representing SMS messages basing on two approaches, namely information gain ratio and chi-square metric. These features are employed as input to two distinct Bayesian classifiers. The authors introduced a real-time mobile application for Android devices based on the proposed framework and evaluated its effectiveness on a large SMS collection. Junaid et al. [15] studied the possibility of using evolutionary classifiers to filter SMS Spam. The authors compared five supervised learning algorithms with four evolutionary classifiers. Experimental evaluation on a real world dataset of SMS has shown that Michigan style classifier outperforms the others in terms of classification accuracy. Almeida et al. [16] compared the Spam detection accuracy of various supervised learning algorithms using a large dataset of SMS. The authors tested two different methods of tokenization and 14 classifiers including 8 variants of Naive Bayes, linear SVM, Minimum Description Length classifier, k-NN, decision tree learner C4.5, and PART (A rule learner). The experimental results have shown that SVM outperforms the other classifiers in terms of Spam detection. Cai et al. [17] proposed a new method to detect unwanted Short Messages using Winnow algorithm [18]. Experimental evaluation on a Chinese Spam dataset has shown the performance of their filter.

## 3. The proposed SMS Spam filter

The proposed filter works in two steps of training and testing.

First step (training process): In this step an improved version of SVDD called Support Vector Domain Description with a small sphere and parametric volume (SSPV-SVDD) [6] will be trained by a collection of SMS that contains two types of messages: Spam and Hams. The goal is to enclose each type of SMS with the smallest hypersphere where the boundary will be used later to detect new Spam. This step begins by applying preprocessing methods to the training dataset since SMS, raw text data, can't be used directly as input to SVDD. The first method is N-grams, the latter transforms a message from text to a new format typically an n-dimensional vector of integers each one represents the frequency of the N-gram in the message. Since the number of N-grams is very large, a second preprocessing technique based on information gain (IG) metric is required. The aim is to reduce the set of N-grams by selecting the most relevant of them. At the end of the preprocessing the collection of SMS will become a set of numeric vectors that will be used as input to SSPV-SVDD to perform the training. The result of this step is two minimal hyperspheres the first encloses the set of Spam while the second surrounds the set of benign SMS.

Second step (testing process): In this step of novelty detection a new unknown short message $X$ will be presented to the filter to decide whether or not the message is a Spam. Firstly the suspect SMS will be transformed to a numeric vector $V$ by applying the same preprocessing used in the training. Then, basing on the two smallest hyperspheres found previously, the Euclidian distance from $V$ to the center of each hypersphere will be evaluated and compared to the radiuses. In principle $X$ will belong to the class represented by the smallest hypersphere in which $V$ exists.

In what follows a detailed description of the proposed filter is presented.

### 3.1. SMS preprocessing using N-grams and IG

N-grams is an overlapping sequences of $N$ characters in a text, although the term can include the notion of any co-occurring set of characters. N-grams was introduced by Shannon [19] to analyze the information content of English text. N-grams has been widely used in many application areas including, text categorization, information retrieval, text compression, etc. To describe a text in terms of N-grams a fixed-size window of $N$ characters is moved over the text, sliding forward by a fixed number of characters (Generally

one character) at a time. At each position of the window the sequence of characters is registered. These sub-strings yield a map to a high-dimensional vector space, where each dimension is associated with the occurrence of one N-gram. Although N-grams provides generic and effective means to model data, the exponential growth of the resulting vector space will raise drastically the computational cost and the storage space required for further processing. To reduce efficiently the size of the vector space a criterion called Information Gain will be used. Generally the evaluation of IG is performed as follows:

Suppose we are given a set $S$ of $M$ classes that contains $s$ labeled training data points where each class $I$ comprises $s_i$ points. Expected information needed to classify a given sample is evaluated as follows:

$$I(s_1, s_2, \dots, s_M) = -\sum_{i=1}^{M} \frac{s_i}{s} log_2\left(\frac{s_i}{s}\right) \quad (1)$$

An attribute $A$ with values $\{A_1, A_2, .., A_v\}$ can divide the set $S$ into $v$ subsets $\{S_1, S_2, .., S_v\}$ where $S_j$ is the subset having the value $A_j$ for attribute $A$ and contains $s_{ij}$ points of class $i$. The entropy of an attribute $A$ can be written as follows:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+..+s_{Mj}}{s} \times I\left(s_{1j}, s_{2j}, \dots, s_{Mj}\right) \quad (2)$$

Finally, the information gain of $A$ can be evaluated using the following equation:

$$Gain(A) = I(s_1, s_2, \dots, s_M) - E(A) \quad (3)$$

After evaluating and sorting all of the attributes of $S$ according to the descending order of their gain. User can choose the first $K$ attributes as the most influencing. $K$ is supposed to be the best compromise between accuracy and processing speed.

### 3.2. SMS Spam filtering using SSPV-SVDD

Support Vector Domain Description (SVDD) is a supervised machine learning method developed first by Tax and Duin [21, 22]. Inspired by the idea of Support Vector Machines (SVM) [8, 9], SVDD describes a given dataset with a minimal hypersphere that encloses all or most of target data. The hyper-spherical boundary will be used to detect novel data or outliers. SVDD classifier has many advantages including: It has a solid mathematical foundation based on the statistical learning theory. Also, it uses kernel functions that can map an irregularly shaped data points represented in the input space into a high dimensional feature space in which a hyper-spherical boundary around the data points can be found. The mapping is performed simply by replacing all inner products in the mathematical formulations of SVDD with a properly chosen kernel function. In addition, training a given dataset with SVDD implies solving a constrained quadratic problem (QP) with a single minimum which avoids the risk of becoming trapped by local minimums. Moreover the classification of a new sample requires checking a decision function basing only on a small subset of the training data known as support vectors (SVs) which accelerates the classification process. Furthermore training SVDD requires setting the values of a small set of parameters which facilitates the use of this classifier. The success of SVDD has been proved in a variety of novelty detection applications, such as network intrusion detection [10], Email Spam filtering [11], malware detection [12], etc .

Among different new extensions of SVDD, we have been particularly attracted by a new version this classifier called SSPV-SVDD [6] that has shown good generalization ability. The main difference between the standard SVDD and SSPV-SVDD is that this last has introduced a new parameter on SVDD that offers the following advantages: i) It permits to customize the hyperspherical boundary between different classes ii) It represents a compromise between the presence of false positives and false negatives iii) It permits to distinguish between the set of samples existing on the hyper-spherical boundaries.

As SVDD, SSPV-SVDD considers a dataset $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ with $i = 1, .., N$ and $x_i \in \mathbb{R}^d$ .The label $y_i$ equals $+1$ for the target objects, and $-1$ for the negative ones. The objective of SSPV-SVDD is to find the smallest hypersphere, with a center $a$ and a radius $R$ that includes the most of target samples and excludes the majority of negative ones following the value of a regularization parameter called $p$ . This problem can be formulated mathematically as follows:

Minimize:

$$R^2 + C\sum_{i=1}^{N} \varepsilon_i \quad (4)$$

Subject to:

$$\|x_i - a\|^2 \leq R^2 - p.y_i + \varepsilon_i \;\; \forall\, i = 1, .., N \;\; with \;\; y_i = +1$$

$$\|x_i - a\|^2 \geq R^2 - p.y_i - \varepsilon_i \;\; \forall i = 1, .., N \;\; with \;\; y_i = -1$$

Where $\|.\|$ is the Euclidean norm, $\varepsilon_i$ are slack variables that measure the amount of violation of the constraints, $p$ is a strictly positive real number. $C$ is a positive parameter that allows the presence of errors. It gives the tradeoff between the volume of the minimal sphere and the respect of the inequality constraints.

It's an optimization problem with constraints that may be solved through Lagrange multipliers. The primal problem of SSPV-SVDD can be expressed as follows:

$$L(R, \varepsilon, a) = R^2 + C\sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \alpha_i y_i (R^2 - \|x_i - a\|^2 - p.y_i) - \sum_{i=1}^{N} \varepsilon_i \mu_i \quad (5)$$

$\mu_i$ and $\alpha_i$ are Lagrange multipliers. The annulation of the partial derivatives of $L$ with respect to $R$, $a$, $\varepsilon_i$ gives:

$$\frac{\partial L}{\partial R} = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 1 \quad (6)$$

$$\frac{\partial L}{\partial a} = 0 \Rightarrow a = \sum_{i=1}^{N} \alpha_i x_i y_i \quad (7)$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \Rightarrow \alpha_i = C - \mu_i \quad (8)$$

Thus, the dual optimization problem can be written as:
Maximize:

$$L(\alpha) = -\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j + \sum_{i=1}^{N}\alpha_i(y_i x_i x_i + p) \quad (9)$$

Subject to :

$$0 \le \alpha_i \le C \quad and \quad \sum_{i=1}^{N}\alpha_i y_i = 1 \quad (10)$$

The maximization of the eq (9) under the constraints described by eq (10) yields the smallest sphere with the center $a$ and the radius $R$ that encloses only the target objects. The center of that sphere can be evaluated through eq (7) and the radius $R$ can be expressed as follows:

$$R^2 = \|x_s - a\|^2 + y_s.p \quad (11)$$

Where $x_s$ is a training point that belongs to the set of Support Vectors having $0 < \alpha_s < C$.

To classify a new unknown sample z the Euclidian distance from the latter to the center $a$ must be evaluated and compared with $R$. If $R > \|z - a\|$ then we conclude that $z$ belongs to the target class. In M-class problems each class $l$ with $1 \le l \le M$ is represented by a minimal sphere with a center $a_l$ and a radius $R_l$ found by applying SSPV-SVDD on the $l^{th}$ class. In this case the decision function that gives the class of $z$ can be expressed as follows [20]:

$$argmax_{1\le l\le M}\left(1 - \sqrt{\frac{\|z - a_l\|^2}{{R_l}^2}}\right) with \ \|z - a_l\|^2 < {R_l}^2 \quad (12)$$

As inherited from SVDD, SSPV-SVDD can be directly and implicitly extended to kernel SSPV-SVDD by replacing inner products in the equations above with an appropriate kernel function. In this work, Gaussian kernel function was employed since it's usually used in pattern recognition problems and it achieves good classification results.

## 4. Experimental evaluation

### 4.1. Dataset description and experimental setting

In order to test practically the accuracy of the proposed filter in terms of SMS Spam detection, we propose three experiments using a dataset of 5574 SMS written in English and tagged according being ham or Spam [4]. This collection was widely used as benchmark dataset in SMS Spam researches. Table 1 presents the basic statistics of this database and Figure 1 shows four samples of ham and Spam SMS taken from this dataset. In each experiment, generally one of the filter parameters, most influencing, will be varied while the others are kept constant in order to investigate its role. The parameters concerned are $N$ in N-grams method, $K$ in Information gain metric and $p$ in SSPV-SVDD classifier. The values of the Gaussian width $\sigma$ are chosen in the interval $\{0, 0.5, .., 5\}$ and the regularization parameter $C$ is fixed at $100$. To perform the experiments the whole dataset of SMS will be Split randomly into two disjoints subsets: The first contains 80% of Spam and 80% of Hams SMS and the second contains 20% of Spam and 20% of Hams SMS. As explained before the proposed filter works in two steps the training will be performed with the first subset and the testing will be performed with the second.

**Table 1**. Basic statistic of the SMS spam dataset

| Type of the SMS | Number of messages | Percentage |
|---|---|---|
| Hams | 4827 | 86.6 % |
| Spams | 747 | 13.4 % |
| Total | 5574 | 100% |

**spam** 100 dating service cal; l 09064012103 box334sk38ch
**spam** You have won ?1,000 cash or a ?2,000 prize! To claim, call09050000327
**ham** I'll be late...
**ham** Waiting for your call.

**Fig 1**. Samples taken from the SMS Spam dataset

### 4.2. Numerical results

✓ Effectiveness of the parameter $p$ in SSPV-SVDD on SMS Spam filtering

The objective of this experiment is to compare the SMS classification accuracy (SCA) of standard SVDD with SSPV-SVDD using different values of $p$. To perform this comparison N-grams method was applied with a fixed value of $N = 4$. Then the most significant N-grams was selected using IG metric with $K = 500$. Next standard SVDD and SSPV-SVDD with $p = 10^{-6}, 10^{-5}$ was trained by the training subset and tested by the testing subset. Figure 2 shows the variation of SCA with different values of σ. The figure can be divided into two parts: In the training process, generally SCA increases with the growth of σ and in the majority of cases SSPV-SVDD outperforms the standard SVDD. This latter gives the lowest result in terms of SCA with a maximum of 92.75%. By applying SSPV-SVDD with $p = 10^{-6}$ the SCA increases and reaches a maximum of 94.39%. By increasing the value of $p$ to $10^{-5}$ the SCA increases too and achieves a maximum of 95.13%. In the testing process, the same previous remarks can be made concerning the variation of the SCA and the superiority of SSPV-SVDD over standard SVDD. Also, the latter gives the lowest SCA with a maximum of 88.41%. By applying SSPV-SVDD with $p = 10^{-6}$ the SCA grows and reaches a maximum of 89.23%. By increasing the value of $p$ to $10^{-5}$ the SCA increases too and achieves a maximum of 89.32%. In conclusion, the experimental results have justified the reason behind choosing SSPV-SVDD as classifier in the proposed filter instead of standard SVDD.

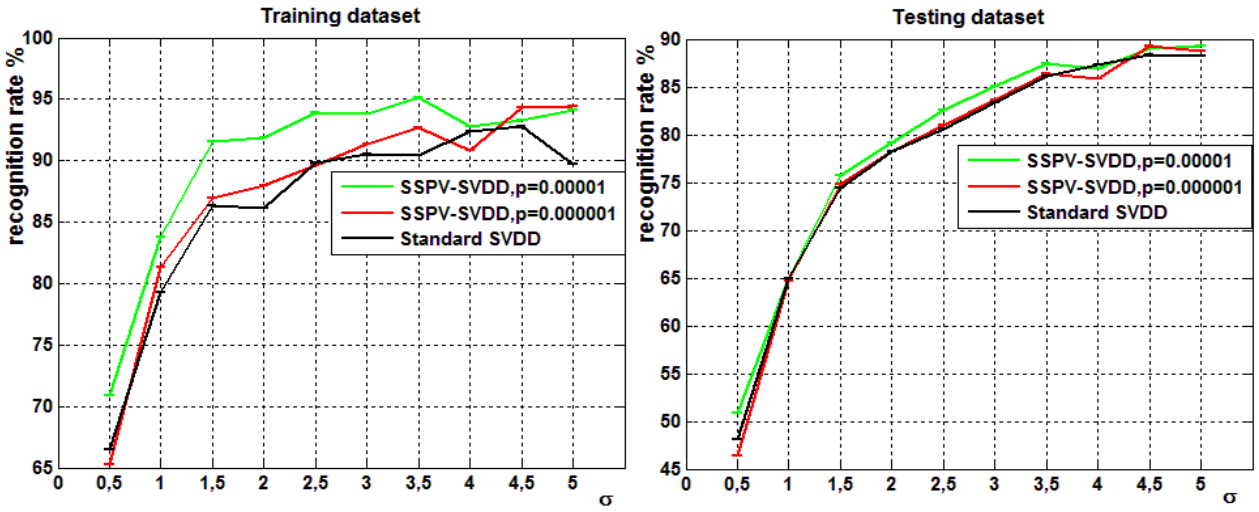


**Fig 2**: SMS classification accuracy of training and testing datasets using Standard SVDD and SSPV-SVDD with N = 4 and K = 500

✓ Effectiveness of the parameter $N$ in N-Gram on SMS Spam filtering

The goal of this experiment is to investigate the role of the parameter $N$ in N-grams on SMS classification accuracy. To perform this test, N-grams method was applied with $N = 2,4$ and 6 then the most significant N-grams was selected using IG measurement with $K = 500$, next the classifier SSPV-SVDD with $p = 10^{-5}$ (The best value of $p$ found in the previous experiment) was trained by the training subset and tested by the testing subset. Figure 3 shows the variation of SCA with different values of $\sigma$ and $N$. Also the figure can be divided into two parts: The training process, when the parameter $N$ is equal to 2 or 4, SCA increases with the growth of σ and gives good performance. In the other hand, when $N$ reaches the value 6 the SCA decreases drastically and yields inconsistent and very poor results. Numerically, with $N = 2$ the SCA reaches to a maximum of 88.81%. By rising the value of $N$ to 4 the proposed filter becomes more accurate and its SCA grows to 95.13%. By increasing again the value of $N$ to 6 the SCA gives, unexpectedly, very poor and inconsistent results with a maximum of 20.36%. In testing process, the same global observation about the variation of SCA can be made for the three values of $N$. Numerically, with $N = 2$ SCA gives a maximum of 68.22%. By increasing the value of $N$ to 4 the SCA yields significant improvement and reaches a maximum of 89.32%. By increasing again the value of $N$ to 6 the SCA gives, unexpectedly, very poor and unstable results with a maximum of 17.68%. The decrease of the performance observed when $N = 6$ can be explained by the fact that N-grams with a large-size window can lose an important number of significant words. Because SMS are short and their text is generally expressed with idioms and abbreviations. The main conclusion that can be drawn from the results of this experiment is that the parameter $N$ in N-grams plays a crucial role in the proposed filter and must be chosen properly according to the problem under investigation.

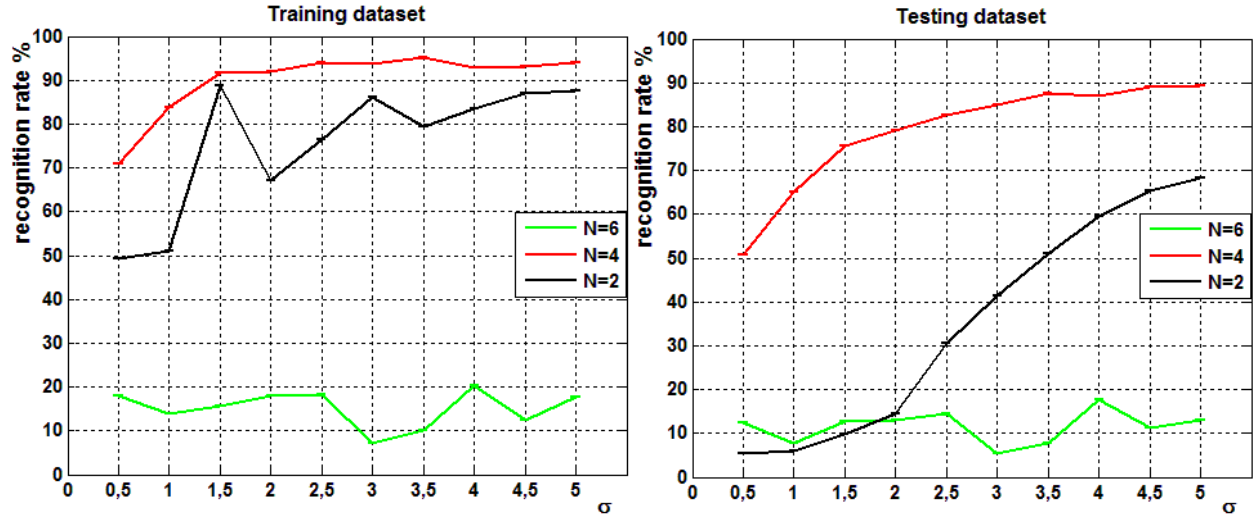


**Fig. 3**. SMS classification accuracy of training and testing datasets using SSPV-SVDD with $N = 2,4,6$, $K = 500$ and $p = 10^{-5}$

✓ *Effectiveness of the parameter K in IG metric* on SMS Spam filtering

The aim of this experiment is to study the role of the parameter $K$ that represents the size of the most significant N-grams evaluated with IG on SCA. To perform this test N-

grams method was applied first with a fixed value of $N = 4$ (The best value found in the second experiment). Then, the most significant N-grams was selected using IG metric with different values of $K$ namely $100, 300$ and $500$. Next, the classifier SSPV-SVDD with $p = 10^{-5}$ (The best value in the first experiment) was trained by the training subset and tested by the testing subset. Figure 4 describes the variation of the SCA with different values of σ and $K$. Again, the figure can be divided into two parts: The training process: Generally when the parameter $K$ is equal to 100 the proposed filter gives very poor performance in terms of SCA whatever the value of σ. By contrary when $K$ takes the value 300 or 500 the proposed filter gives good SCA. Numerically, with $K = 100$ SCA reaches to a maximum of 35.63%. By rising the value of $K$ to 300 the proposed filter becomes more accurate and its SCA grows to 94.30%. By increasing again the value of $K$ to 500 SCA increases too and reaches 95.13%. In the testing process, the same global observation about the variation of SCA with $K$ can be made. Numerically, with $K = 100$ SCA gives a maximum of 36.17%. By increasing the value of $K$ to 300 SCA yields significant improvement and reaches a maximum of 89.21%. By increasing again the value of $N$ to 500 SCA gives a maximum of 89.32%. The good performance observed in the last two tests can be explained by the fact that when $K$ (The size of the most significant N-grams) increases the representation of the SMS dataset becomes more detailed and by consequence SSPV-SVDD can describe accurately the SMSs. The main conclusion that can be extracted from this experiment is that the parameter $K$ in IG metric plays an important role in the proposed filter and must be chosen as large as possible (Taking into consideration the space and time complexities required by SSPV-SVDD classifier).

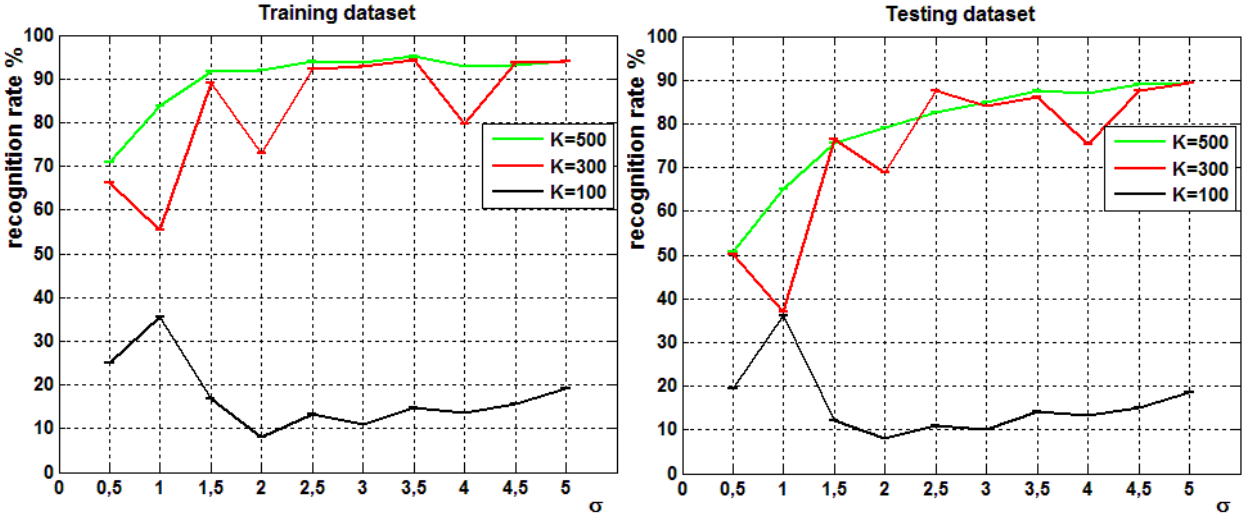


**Fig. 4.** SMS classification accuracy of the training and testing datasets using SSPV-SVDD with $N = 4$, $K = 100, 300, 500$ and $p = 10^{-5}$

## 5. Conclusion

In this paper we presented a new SMS Spam filter based on three components: N-grams method for features extraction, information gain metric for significant features selection and an improved version of standard SVDD named SSPV-SVDD for SMS classification. In order to investigate the accuracy of the proposed filter in terms of real SMS Spam detection, three experiments were performed on a large dataset of SMS containing 5574 short messages. The first experiment is interested in comparing the classification performance of standard SVDD and SSPV-SVDD with different values of the regularization parameter $p$. Experimental results have shown that SSPV-SVDD outperforms the standard SVDD in terms of SMS classification accuracy especially with a good choice of the value of $p$. The objective of the second experiment is to study the effect of the size of the window represented by $N$ in N-gram method on the filter accuracy. Experimental results have shown that the parameter $N$ plays an important role and must be chosen properly. The third experiment has as objective to test the effect of the parameter $K$ that represents the number of the most relevant N-grams selected with IG metric on SCA. Experimental results have proved that as the parameter $K$ grows the classification accuracy of the proposed filter increases and becomes more stable.

Empirically, the optimal values of the filter parameters: $p$, $N$ and $K$ are respectively $0.00001$, $4$ and $500$. With those values the proposed filter shows promising SCA rate that reaches 95.13% in the training process and 89.32% in the testing.



## References

1. J.M. Gomez Hidalgo, G.C. Bringas, E.P. Sanz and F.C. Garcia, Content based SMS spam filtering, In Proceeding of the ACM symposium on document engineering, Amsterdam, The Netherlands, pp. 107-114 (2006).
2. M. Healy, S.J Delany and A. Zamolotskikh, An assessment of case base reasoning for short text message classification, In: Norman Creaney (ed.) Proceedings of the 15th Irish Conference on Artificial Intelligence & cognitive Science (AICS'05), pp. 257-266 (2005).
3. M. Taufiq Nuruzzaman, C. Lee, M. F. A. b. Abdullah, and D. Choi, Simple SMS spam filtering on independent mobile phone,

Security and Communication Networks, Vol. 5 (10), pp. 1209-1220 (2012).
4. UCI repository of machine learning databases, http://archive.ics.uci.edu/ml/.
5. D. Longzhen, L. Nan and H. Longjun. A New Spam Short Message Classification. In Proceedings of the First International Workshop on Education Technology and Computer Science,Vol. 2, pp. 168-171 (2009).
6. M. El boujnouni, M. Jedra, N. Zahid, A small sphere and parametric volume for support vector domain description, Journal of Theoretical and Applied Information Technology, Vol. 46(1), pp. 471-478 (2012).
7. T.M Mahmoud and A.M Mahfouz, SMS Spam Filtering Technique Based on Artificial Immune System, International Journal of Computer Science Issues, Vol. 9, Issue 2, N° 1, pp. 589-597 (2012).
8. V. Vapnik, Statistical Learning Theory, Wiley, NY (1998).
9. V.Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, NY (1995).
10. I. Kang, M.K. Jeong and D. Kong, A differentiated one-class classification method with applications to intrusion detection, Expert Systems with Applications, Vol. 39(4), pp. 3899-3905 (2012).
11. M. El boujnouni, M. Jedra, and N. Zahid. Email Spam filtering using the combination of two improved versions of Support Vector Domain Description, In Proceedings of the 10th International Conference on Computational Intelligence in Security for Information Systems (CISIS15), International Joint Conference Advances in Intelligent Systems and Computing, Vol. 369, Burgos, Spain, pp. 99-109, 15-17 June (2015).
12. M. El boujnouni, M. Jedra, and N. Zahid, New malware detection framework based on N-grams and SVDD with SMO, Journal of Information Assurance and Security, Vol. 11(4), pp. 223-232 (2016).
13. I. Joe and H. Shim, An SMS spam filtering system using support vector machine, Lecture Notes in Computer Science, Vol. 6485, pp. 577- 584 (2010).
14. A.K Uysal, S. Gunal, S. Ergin and E.S Gunal, A novel framework for SMS spam filtering, In Proceedings of International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Trabzon, Turkey, 2-4 July (2012).
15. M.B. Junaid and M. Farooq, Using Evolutionary Learning Classifiers to do Mobile Spam (SMS) Filtering. In Proceedings of the 13th annual conference on Genetic and evolutionary computation, Dublin, Ireland, pp. 1795-1802, July 12-16 (2011).
16. T.A. Almeida, J.M.G. Hidalgo and A. Yamakami, Contributions to the study of SMS Spam Filtering: New Collection and Results. In Proceedings of the 11th ACM Symposium on Document Engineering, CA, USA, pp. 259-262, Sept 19-22 (2011).
17. J. Cai, Y. Tang and R. Hu. Spam Filter for Short Messages Using Winnow. In Proceedings of the International Conference on Advanced Language Processing and Web Information Technology, Liaoning, China, pp. 454 -459, 23-25 July (2008).
18. N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. Machine Learning, Vol. 2(4), pp. 285-318 (1988).
19. C.E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379-423& 623-656, 1948.
20. M. El Boujnouni, M., Jedra and N. Zahid, New decision function for support vector data description. Journal of Information and Systems Management. Vol. 2(3), pp. 105-115 (2012).
21. DMJ. Tax and RPW. Duin, Data domain description using support vectors, in Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, pp. 251-256 (1999).
22. DMJ. Tax and RPW. Duin, Support vector data description, Machine Learning, Vol. 54, pp. 45-66 (2004).