The 29th International Conference on Systems, Signals and Image Processing, IWSSIP 2022, Sofia, Bulgaria, June 1-3, 2022 (Accepted version)

Semi-supervised novelty detection with one class SVM for SMS spam detection

Suleiman Y. Yerima [1] and Abul Bashar [2]

[1] Cyber Technology Institute
Faculty of Computing, Engineering and Media,
De Montfort University, Leicester, United Kingdom
syerima@dmu.ac.uk

[2] College of Computer Engineering & Sciences,
Prince Mohammad Bin Fahd University, P.O. Box 1664, Al-Khobar 31952, Saudi Arabia
abashar@pmu.edu.sa

***Abstract*— The volume of SMS messages sent on a daily basis globally has continued to grow significantly over the past years. Hence, mobile phones are becoming increasingly vulnerable to SMS spam messages, thereby exposing users to the risk of fraud and theft of personal data. Filtering of messages to detect and eliminate SMS spam is now a critical functionality for which different types of machine learning approaches are still being explored. In this paper, we propose a system for detecting SMS spam using a semi-supervised novelty detection approach based on one class SVM classifier. The system is built as an anomaly detector that learns only from normal SMS messages thus enabling detection models to be implemented in the absence of labelled SMS spam training examples. We evaluated our proposed system using a benchmark dataset consisting of 747 SMS spam and 4827 non-spam messages. The results show that our proposed method outperformed the traditional supervised machine learning approaches based on binary, frequency or TF-IDF bag-of-words. The overall accuracy was 98% with 100% SMS spam detection rate and only around 3% false positive rate.**

Keywords—SMS; spam detection; smishing; semi-supervised learning; novelty detection model; One Class Support Vector Machine

I. Introduction

Short Messaging Service (SMS) is a popular text messaging service that allows fixed line or mobile phone devices to exchange short messages using standard communication protocols. The volume of SMS text messages has grown steadily over the past decade as mobile phone usage continues to increase. According to the latest data from GSMA intelligence, there are 5.31 billion unique mobile phone users around the world today and this grew by 95 million in the past year [1]. About 5 billion people globally, i.e. 65% of the world's population send and receive SMS messages [2][3].

Unfortunately, the popularity of SMS has also led to an increase in unwanted and unsolicited messages known as spam. This leaves millions of mobile subscribers more exposed to fraud where the intention may be to gain personal data to sell for profit, or to trick recipients into clicking on a malicious URL, etc. Hence, the detection and elimination of SMS spam is a critical task necessary to protect millions of subscribers on mobile networks around the world.

In the last few years, various approaches have been proposed for detection of spam, particularly for email. However, there has recently been an increase in studies that have focused on mobile SMS spam detection with several studies investigating machine learning based models for SMS spam classification. However, most of these studies have been based on supervised learning approach. Supervised machine learning approaches demands a large amount of labelled data which is not always available in real applications [4].

In this paper, we propose a system for SMS spam detection based on a semi-supervised novelty detection approach. Our system adopts an anomaly detection approach where normal non-spam SMS text messages are used to build a classification model utilizing One Class SVM (OC-SVM). The advantage of this method is that it is based on an anomaly detector that can be built from non-spam SMS messages thus eliminating the need for labelled dataset that is necessary for supervised learning. Furthermore, being a semi-supervised approach, our proposed system is less susceptible to the problems frequently encountered with imbalanced datasets during supervised learning. We performed evaluation experiments using the benchmark dataset containing 747 SMS spam and 4827 non-spam messages and obtained 98% overall accuracy with 100% true positive rate for SMS spam, thereby demonstrating the effectiveness of the proposed approach. The rest of the paper is organized as follows. In section II, we review related work. In section III we discuss our methodology, while in section IV experimental results are presented. Finally, we conclude the paper and outline future work in section V.

II. Related Work

Spam detection is an active area of research with several research works having applied machine learning techniques to the problem. In [5], the authors applied TF-IDF (Term Frequency-Inverse Document Frequency) Bag of Words approach in conjunction with Random Forest (RF) to achieve an accuracy of 97.5%. RF was shown to outperform Decision Tree (DT), Support Vector Machines (SVM), K-Nearest Neighbour (KNN) and Multinomial Naïve Bayes (MNB) classifiers and the study was based on the same UCI SMS spam dataset used in our paper. In [6], an approach was proposed to detect and filter SMS spam messages using machine learning algorithms. They used 10 feature types derived from studying the characteristics of spam messages in-depth (i.e. hand-engineered features). Some of these features include presence of URLs (Uniform Resource Locator), presence of mathematical features, presence of dots, presence of special symbols, presence of mobile numbers, and message length. They performed experiments with 5 machine learning classifiers and achieved 96.5% true positive rate and 1.02% false positive rate with the RF classifier.

In [7], the authors propose an intelligent framework for filtering email and SMS messages, using Dendric Cell Algorithm (DCA), an immune-inspired classification algorithm. The method was shown to outperform KNN, SVM and Naïve Bayes (NB) algorithms and their combination using majority voting. In [8], the authors applied both deep learning and shallow learning models using supervised learning for the detection of SMS spam. The deep learning models i.e. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) outperformed the traditional machine learning classifiers including, Naïve Bayes, Random Forest, Logistic Regression (LR), and Stochastic Gradient Descent classifiers.

In [9], LR, KNN, and DT were used where LR was shown to yield an accuracy of 99% using the 'SMS spam collection dataset' containing 5572 instances (4900 non-spam; 672 spam) from Kaggle repository. In [10], the authors proposed a multimodal architecture based on model fusion where the system is designed to process text and image parts of an email to detect spam. They used Deep Neural Networks (DNN) for spam detection and obtained 98.48% accuracy. In [11], the authors proposed an artificial immune system for spam detection achieving 98.05% accuracy. Popovac et al. [13] proposed a CNN based SMS spam detection model and obtained 91.5% sensitivity (TPR), 99.44% specificity (TNR), 0.955 AUC and 0.938 F1 score respectively using 10-fold Cross Validation (CV). Other deep learning-based works include [14] [15] [16] [17] and [18].

A review of previous works shows that the use of supervised learning is the more popular trend in spam detection. Hence, unlike most previous studies, this paper investigates a semi-supervised novelty detection approach where training is performed on only one class of data i.e., the normal or non-spam data. This makes the trained classifier model an anomaly detector as it provides an outlier detection mechanism that considers the outliers (anomalies) as spam messages. We then evaluate the performance of the model on both spam and non-spam examples, achieving high accuracy results.

III. Methodology

A. Dataset

The dataset used in this study was obtained from [12]. It consists of 5,574 real and non-encoded messages. The messages are labelled as non-spam (ham) or spam. The total number of spam messages contained in the dataset is 747 (13.4%) while non-spam messages are 4827 (86.6%).

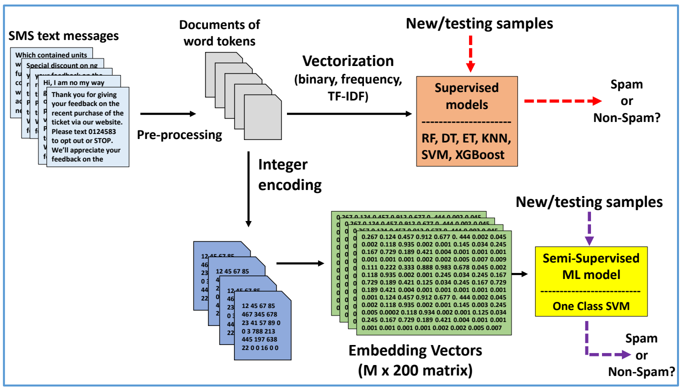

Figure 1: Simplified architecture of the machine learning based spam detection framework used for our study.

B. SMS text pre-processing

The SMS text pre-processing consists of several tasks that were implemented using Python scripts to ultimately create a feature set for training learning models for efficient and high accuracy spam filtering. These tasks are depicted in Figure 1. In our case, the pre-processing steps were mainly aimed at creating input data suitable for training an OC-SVM anomaly detector from non-spam messages. The first pre-processing task was converting all the text in the dataset to lowercase. This was followed by tokenization (splitting the text into individual words). Stopwords (such as 'the', 'and', 'for' etc.) which connect parts of a text were removed, as these were tokens that provide no useful information during the training process. Finally, all the tokens resulting from the pre-processing were saved into a separate document for each SMS message. These documents representing the SMS messages provided the input for the model building process.

C. Building the training features from text documents

The next steps involve tasks to produce features for model training. First, a vocabulary of 10,000 words (i.e. Bag-of-Words - BoW) was produced from the text documents of the training set. Since our model is a semi-supervised novelty detector, the training set consists of only documents of the positive class which is the non-spam messages. From the vocabulary set, we applied integer encoding to convert the words present in each message into integers. If a word appears in the message and does not exist in the vocabulary (BoW), it is represented by a zero. The maximum size retained for each message was 50 and those that consisted of less than 50 words were padded with zeros so that every message was encoded by a set of integers of the same length. The final integer encoded messages were represented as an M x N matrix where N = 50 and M is the total number of non-spam (positive) training messages.

Each integer-encoded word in the M x N matrix is converted into a K-dimensional vector representing the word. K was chosen as 4 to maintain low dimensionality for the feature vectors and enable fast training and prediction. The process of converting the (integer-encoded) words to (K-dimensional) vectors is known as 'embedding' and this was achieved by using an embedding layer from the Keras Python library. To achieve the 4-dimensioanl vector embedding for each word, we created a model (with the Keras embedding layer) that was trained for 50 epochs using the M x N matrix of integer-encoded words as the input to the model and obtaining a resulting M x 200 matrix from the model. The M x 200 matrix is used to train the OC-SVM model with M instances of vectors of length = 200 each representing an SMS message.

D. OC-SVM model implementation

The dataset used in this study is an imbalanced dataset with spam data as the minority class. By treating the majority class as 'normal', anomaly or novelty detection techniques can be used to detect spam as 'novelty' or 'anomaly'. One-class classification involves fitting a model on 'normal' data and predicting whether new data is normal or an anomaly/outlier. We chose the OC-SVM classifier for this study and trained it on the non-spam data. OC-SVM is an unsupervised algorithm that learns a decision function for the purpose of novelty detection with the goal of classifying new data as similar or different to the training set. The main difference from the standard SVM is that it is fit in an unsupervised manner and only provides a hyperparameter *nu* that controls the sensitivity of the support vectors, instead of normal SVM hyperparameters used for tuning the margin. The parameter *nu* should be tuned to approximate the ratio of outliers in the training data. For example, in our experiments we selected *nu* = 0.03 meaning that approximately 3% of the training set will lie outside the boundaries of the support vectors.

IV. EXPERIMENTS AND RESULTS

In our study, we performed several experiments to evaluate the performance of the proposed model. We compared the performance of the semi-supervised approach to supervised methods that were built using the classical BoW approaches after preprocessing of the SMS text, but in this case the vocabulary is built from both spam and non-spam text documents designated as the training set. Every message was converted into a vector-based representation of each of the terms (tokens) in the vocabulary set. Since this leads to sparse vectors of high dimension, we applied chi-squared feature selection algorithm, using this to select the top 200 features. Three different vector representations were used including: binary, frequency and TF-IDF (i.e., Term Frequency-Inverse Document Frequency). Each of these feature vectors were used to train several supervised models including Random Forest (RF), k-nearest neighbours (KNN), eXtreme Gradient Boosting (XGBoost), Support vector machines (SVM) and Extra Trees (ET). The performance of these models was then compared to results of the proposed OC-SVM novelty detector model. The models were implemented using Scikit Learn and other associated Python libraries and the experiments were performed on an Ubuntu Linux 16.04 64-bit Machine with 4 GB RAM. We used 70:30 training-testing split to evaluate all the supervised models. The results are the average of 10 runs with random selection of instances.

A. Results of the supervised learning models with binary feature vactors

Table I shows results from models trained with binary features, i.e., features that capture the presence or absence of the tokenized words in the vocabulary. Feature selection was applied to improve performance. Non-spam true positive rates are all above 99.4% while the highest spam detection rate is 81.6% with supervised linear SVM. Class imbalance in the dataset accounts for the large disparity.

Table I: Results of supervised learning with binary features.

	TPR (S)	TPR (NS)	Accuracy	F1
ET	0.814	0.998	0.972	0.970
KNN	0.55	1.000	0.939	0.931
DT	0.788	0.994	0.966	0.965
SVM	0.816	0.995	0.970	0.969
RF	0.797	0.999	0.970	0.969
XGBoost	0.811	0.996	0.971	0.969

B. Results of the supervised learning models with frequency feature vactors

Table II shows results from models trained with frequency features, i.e., features that capture how often tokenized words in the vocabulary appear in a message. Feature selection was applied to improve performance. The true positive rates for non-spam messages are all above 98.9%, while the highest spam detection rate is 87.1% with the Extra Tree classifier. For majority of the models, frequency features performance is better that binary features performance.

Table II: Supervised learning with frequency features.

	TPR (S)	TPR (NS)	Accuracy	F1
ET	0.871	0.992	0.974	0.973
KNN	0.631	0.999	0.947	0.942
DT	0.780	0.989	0.959	0.958
SVM	0.539	0.993	0.925	0.915
RF	0.838	0.994	0.971	0.970
XGBoost	0.837	0.992	0.971	0.971

C. Results of the supervised learning models with TF-IDF feature vectors

Table III shows results from models trained with TF-IDF features. Feature selection was applied to improve performance. Spam messages had the highest detection rate of 84.7% while the non-spam messages had a detection rate of 99.4% or better. Except for SVM and KNN, the other models had better spam detection performance with the frequency features compared to TF-IDF features. However, for non-spam detection rate the models obtained better results with TF-IDF compared to frequency features.

Table III: Results of supervised learning with TF-IDF features.

	TPR (S)	TPR(NS)	Accuracy	F1
ET	0.817	0.999	0.973	0.972
KNN	0.616	0.999	0.946	0.948
DT	0.789	0.994	0.966	0.965
SVM	0.847	0.995	0.973	0.973
RF	0.797	0.999	0.971	0.969
XGBoost	0.811	0.996	0.971	0.969

D. Results of the supervised learning models vs. proposed OC-SVM based novelty detecton model.

In Table IV, we compare the results of the proposed OC-SVM novelty detection method with the best results from the binary, frequency, and TF-IDF with supervised learning. We can see that OC-SVM outperformed them by recording the best spam detection rate of 100%. This means that the model trained with non-spam extracted data was able to detect all the spam messages. None of the supervised learning models investigated was able to achieve close to this result. The overall accuracy from OC-SVM was 98% which was better than the best binary model (97.2%), best frequency-based model (97.4%) and the best TF-IDF based model (97.3%). Note that the OC-SVM model was built from 70% of the non-spam data and evaluated on the other 30% of the non-spam data plus all of the spam data. This was repeated several times by randomly choosing another set of 70% from the non-spam data for training and the average of 10 runs was computed to achieve the results as shown in Table IV.

Table IV: Novelty detection vs. supervised learning results.

	TPR (S)	TPR (NS)	Accuracy	F1
OC-SVM	**1.000**	**0.968**	**0.980**	**0.980**
Best binary (Extra Trees)	0.814	0.998	0.972	0.970
Best frequency (Extra Trees)	0.871	0.992	0.974	0.973
Best TF-IDF (SVM)	0.847	0.995	0.973	0.973

V. Conclusions and future work

SMS spam detection has been an active area of research with supervised machine learning being the most popular detection approach. This approach is inefficient for imbalanced data. Hence, in this paper we proposed a system for SMS spam detection based on semi-supervised novelty detection using One Class SVM (OC-SVM). It employs only non-spam data for training, making it feasible to implement in the absence of labelled spam data. After pre-processing of non-spam data, integer encoding is applied, followed by a low dimensional vector embedding which produces the input data for the OC-SVM model training. The proposed approach achieved 100% true positive rate for SMS spam detection and an overall accuracy of 98% which was better than the results of all the 18 bag-of-words based supervised machine learning models that we implemented and evaluated on the same dataset. For future work, we aim to investigate other types of novelty detection models based on semi-supervised machine learning.

References

[1] GSMA intelligence available online: https://www.gsmaintelligence.com/ [Accessed 30 March, 2022]

[2] Datareportal "Digital around the world" Available online: https://datareportal.com/global-digital-overview [Accessed 30 March, 2022]

[3] Slicktext "44 Mind-Blowing SMS Marketing and Texting Statistics" online: https://www.slicktext.com/blog/2018/11/44-mind-blowing-sms-marketing-and-texting-statistics/ [Accessed 30 March, 2022]

[4] I. Ahmed, R. Ali, D. Guan, Y.-K. Lee, S. Lee, and T. Chung, "Semi-supervised learning using frequent itemset and ensemble learning for SMS classification," Expert Systems with Applications, vol. 42(3), 2015, pp. 1065-1073.

[5] N. N. A. Sjarif, N. F. M. Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, "SMS Spam Message Detection using Term Frequenct-Inverse Document Frequency and Random Forest Algorithm," in Fifth Information Systems International Conference 2019, Procedia Computer Science 161 (2019) 509–515, ScienceDirect.

[6] N. Choudhary and A. K. Jain. "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique," in Advanced Informatics for Computing Research vol. 712 pp.18-30, 2017.

[7] E. S. M. El-Alfy, and A. A. AlHasan, "Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm," Future Gen. Comput. Syst. vol. 64, pp. 98–107, 2016. doi:10.1016/j.future.2016.02.018

[8] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," Future Gen. Computer Systems, vol. 102, pp. 524–533, 2020.

[9] L. GuangJun, S. Nazir, H. U. Khan and A. Ul Haq, "Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms," Security and Communication Networks, vol. 2020, https://doi.org/10.1155/2020/8873639

[10] H. Yang, Q. Liu, S. Zhou, and Y. Luo, "A spam filtering method based on multi-modal fusion," Applied Sciences, vol. 9, no. 6, p. 1152, 2019.

[11] A. J. Saleh, A. Karim, B. Shanmugam, S. Azam, K. Kannoorpatti, M. Jonkman and F. De Boer, "An intelligent spam detection model based on artificial immune system," Information, vol. 10, no. 6, p. 209, 2019.

[12] Kaggle "SMS Spam collection dataset" Available online: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset [Accessed 31 March 2022]

[13] M. Popovac, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Convolutional Neural Network Based SMS Spam Detection," 26th Telecommunications forum TELFOR 2018. Belgrade, Serbia.

[14] S. Annareddy and S. Tammina "A Comparative Study of Deep Learning Methods for Spam Detection," Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, AnalyticsandCloud) (I-SMAC 2019). 12-14 December, Palladam, India, 66-72.

[15] T. Huang, "A CNN Model for SMS Spam Detection," 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). 25-27 October, Hohhot, China, 851.

[16] S. Gadde, A. Lakshmanarao and S. Satyanarayana, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 358-362, 2021.

[17] P. Poomka, W. Pongsena, N. Kerdprasop, and K. Kerdprasop, "SMS Spam Detection Based on Long Short-Term Memory and Gated Recurrent Unit," International Journal of Future Computer and Communication 8(1), pp. 11-15, 2019.

[18] H. Raj, Y. Weihong, S. K. Banbhrani, and S. P. Dino, "LSTM Based Short Message Service (SMS) Modeling for Spam Classification," In Proc. of the 2018 Int. Conf. on Machine Learning Technologies (ICMLT '18). New York, NY, USA, pp. 76–80, 2018.