International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

IJIEMR Transactions, online available on 21st Nov 2021. Link :http://www.ijiemr.org/downloads.php?vol=Volume-10&issue=Issue 11

10.48047/IJIEMR/V10/ISSUE 11/51

Title **A NOVEL APPROACH FOR SUPERVISED MACHINE LEARNING BASED SMS SPAM DETECTION**

Volume 10, ISSUE 11, Pages: 316-322

Paper Authors **P. ANANDI, CH. HARISH, D. VENKATESH, B. SHANMUKH KUMAR, G.M. VISHNU VARDHAN, A. SASIKANTH**

International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

A NOVEL APPROACH FOR SUPERVISED MACHINE LEARNING BASED SMS SPAM DETECTION

[1] P. ANANDI, [2]CH. HARISH, [3]D. VENKATESH, [4]B. SHANMUKH KUMAR, [5]G.M. VISHNU VARDHAN, [6]A. SASIKANTH

[1]Associate Professor, Department of ECE, Sree Venkateswara College of Engineering, Northrajupalem(VI), Kodavaluru(M), Nellore (DT), Andhra Pradesh, India.

[2,3,4,5,6]B.Tech Scholars, Department of ECE, Sree Venkateswara College of Engineering, Northrajupalem(VI), Kodavaluru(M), Nellore (DT), Andhra Pradesh, India.

ABSTRACT: The Short Message Service (SMS) has been widely used as a communication tool over the past few decades as the popularity of mobile phone and mobile network grows. It allows for the broad spread of "spams," i.e., inferiority news with deliberately false information. The widespread spread of spams has the potential for very negative impacts on people and society. Machine Learning methods for anti-spam filters have been noticeably effective in categorizing spam messages. This paper presents, a novel approach for Supervised Machine Learning based SMS Spam Detection. Dataset used in this research is known as Tiago's dataset. Crucial step in the experiment was data preprocessing, which involved reducing text to lower case, tokenization, removing stopwords. The SMS dataset used was imbalanced, and to solve this problem, we used oversampling and under-sampling techniques. The support vector Machine (SVM), Naïve Bayes (NB), and Logistics Regression (LR) are applied on the spam and ham SMS dataset. SMS spam filter inherits much functionality from E-mail Spam Filtering. Comparing the performance of various supervised learning algorithms we find the support vector machine algorithm gives us the most accurate result.

KEYWORDS: Short Message Service (SMS), Spam Detection, Machine Learning,

I. INTRODUCTION

Technology gradually makes improvements in our daily life, and its development technology is a continuous and ongoing process. As a result, humans depend on technology to perform their tasks in every field of life, like the medical domain, information technology, and communication domain. In all the social media like face book, twitter, bingo etc. spreading the news based on the customer analysis only [1].
Individuals tend to share their data on different sites, though that data is imparted to different organizations that spam individuals to offer their services.

The Short Messaging Service (SMS), commonly referred to as "text messaging" is a service for transmitting short length messages of around 160 characters to different devices such as cellular phones, smartphones and PDAs using standardized communications protocols [2]. SMS is used as an alternate for voice calls in positions where voice communication is either not possible or not desired between the end phone users. It is one of the most flourishing phone service engendering millions of dollars in perquisite for mobile operators yearly. Today's estimates signify that billions of SMS's are sent per day.

Spams are undesirable and unwelcomed messages which are sent electronically. These messages are sent by spammers for different ill wills of taking a hold over user's personal data or tricking them into the subscription of their premium tariff facilities. SMS Spamming in extremely disappointing for the clients: numerous critical and valuable messages can get lost because of spam messages, Spam messages are additionally used to trap individuals, or bait them into purchasing services. As overall utilization of cell phones has grown, another road for e-junk mail has been opened for notorious advertisers. These publicists use instant messages (SMS) to target probable purchasers with undesirable publicizing known as SMS spam [3]. Such messages have the capability of imposing the same threats or even more dangerous aftereffects as of Email spams. In the zone of Email, though spam is a properly handled obstacle but SMS spams are increasing at a high rate of more than 500% in an year.

As time passes, the advertisement of anything is becoming the key feature to improve the business. Due to the importance of advertisement, different advertising agencies now use SMS as a medium of advertisement. This is the fastest way to communicate the business brochure to a typical person. That is why spammer also uses SMS technology to contact people, a source of income for the spammer. There are many tools available that prevent spam SMS, but according to an estimate, a person daily receives the bulk of SMS, and more than 50% are spam SMS. There would be a system that can identify ham, and spam SMS accurately.

Machine learning is one of the most popular topics in the last few decades, and there are a great number of machine learning based classification applications in multiple research areas. Specifically, spam detection is a relatively mature research topic with several established methods. However, most of the machine learning based classifiers were dependent on the handcrafted features extracted from the training data [4].

This study proposed an approach to classify spam and ham SMS using supervised machine learning algorithms. The dataset contains both spam and ham SMS. Firstly, this research applies to preprocessing techniques to clean SMS that include stemming, stop words removal technique, convert to lowercase, punctuation removal, and numeric removal techniques. Oversampling and under-sampling are also performed to get more accurate results.

Spam messages can be classified as redundant messages sent to large number of people at once. The rise of spam messages are based on the following factors: 1) The accessibility to cheap bulk SMS-plans; 2) dependability (since the message comes to the cell phone client); 3) low possibility of accepting reactions from some unaware recipients; and 4) the message can be customized.5) Free services.

II. LITERATURE SURVEY

A. K. Jain and B. B. Gupta, et. al. [5] proposed a method to apply rulebased models on the SMS spam detection problem. The authors extracted 9 rules and implemented Decision Tree (DT), RIPPER, and PRISM to identify the spam messages. we have used text normalization to convert them into standard form to obtain better rules. The performance of the proposed approach is evaluated, and it achieved more than 99% true negative rate. Furthermore, the proposed approach is very efficient for the detection of the zero hour attack too. P. Sethi, V. Bhandari and B. Kohli, et. al. [6], presents SMS spam detection and comparison of various machine learning algorithms. The research was conducted on the Tiago's dataset. In the experiment, Naïve Bayes outperformed Random Forest algorithm and Logistic Regression algorithm. NB provided the results of almost 98.5% accuracy.

N. K. Nagwani and A. Sharaff, et. al. [7] presents SMS spam filtering and thread identification using bi-level text classification and clustering techniques. Besides pre-processing and classification with various classifiers, the experiment included Clustering using K-Means algorithm or NMF Model. After the mentioned steps, a solution for SMS Thread Identification was proposed. The research led to the conclusion that the SVM algorithm performs better in categorizing the SMS messages and the combination of NMF and SVM algorithm gave good results in thread identification.

D. Delvia Arifin, Shaufiah and M. A. Bijaksana, et. al. [8], SMS Spam Corpus and SMS Spam Collection datasets were used, individually and merged. Both of the two used methods, Naive Bayes classifier and FP-Growth Algorithm, accomplished an accuracy rate superior than 90%. The accuracy best average (98.5%) was obtained with the implementation of the FP-Growth algorithm on Tiago's dataset. T. A. Almeida, J. M. G. Hidalgo and T. P Silva, et al. [9] showcased the particulars of a new authentic, open and non-encoded SMS spam compilation which constitutes of maximum number of messages. It is composed of 4,827 mobile ham messages and 747 mobile spams. Furthermore, the authors performed several established machine learning algorithms on their dataset and they came to the conclusion that according to them SVM is a better approach for advance evaluation.

M. Taufiq Nuruzzaman, C. Lee, M. F. A. bin Abdullah and D. Choi, et. al. [10] looked over the efficiency of sieving message spam on independent cellular phones using Text Classification approaches. On an independent mobile various processing were done related to training, filtering, and updating. Their established outcomes display that the projected model was successful in distilling messages hams and spam with moderate efficiency, consuming less memory, and appropriate time was consumed while functioning without taking the help from a machine**.**

III. SUPERVISED MACHINE LEARNING BASED SMS SPAM DETECTION

The architecture of a novel approach for Supervised Machine Learning based SMS Spam Detection is represented in below Fig. 1.

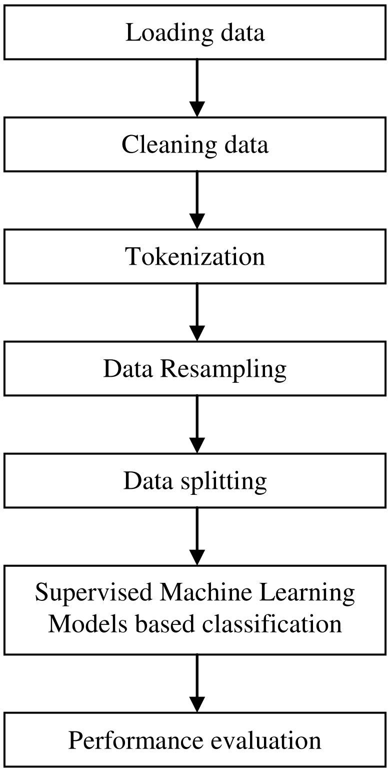

Fig. 1: ARCHITECTURE OF SMS SPAM DETECTION

The dataset used in this research is also known as Tiago's dataset. It is composed of

5,574 English, real and non-encoded messages labeled as ham (non-spam) or spam.

Cleaning of data is performed to improve the training model's learning process [24]. SMS contains stop words, punctuation, and upper and lower case words that can affect and reduce the learning of the training model. The processing is applied after collecting the dataset with an equal number of SMS. Converting text into something an algorithm can work with represents a complex process.
Remove Punctuation: Using punctuation in the text helps the reader to understand the message that is being conveyed clearly. But these marks have no meaning, so they are not helpful for model training, and we remove them in preprocessing.
Remove Numbers: SMS can contain numbers that are not useful in machine learning models training. We remove these numbers to reduce complexity in the features set. We remove these numbers using regular expressions.
Stemming: To convert each word into its root form we used the stemming technique. We used the Porter stemmer technique to perform stemming.
Remove Stopwords: Stopwords are the parts of text but have no meaning, so to focus on the meaningful words during training, we remove stopwords. We remove stopwords using the natural language toolkit.

In the next step, tokenization (splitting text into individual words) was conducted in order to perform stemming. By reducing words to a root, sentiment of the text was preserved. Also, words which serve only for connecting parts of given text (messages), rather than influencing model (so-called stopwords), were removed. Further, extraction of polarity and subjectivity for every message was provided in order to interpret sentiment analysis.

We used Oversampling (SMOTE) and Undersampling Techniques such (Random Undersampling). Over-sampling is a technique where the number of the minority class in the majority class ratio is raised. Oversampling increases the sample size, generating additional features for model training and enhancing the model's accuracy. Random Undersampling technique works by rejecting randomly selected examples of the majority class and deleting them so that the distribution of target classes can be balanced. The data was then prepared by dividing the dataset into training and testing datasets, with 75% of the messages used as the training dataset and 25% was used as the testing dataset.

The support vector Machine (SVM), Naïve Bayes (NB), and Logistics Regression (LR) are some supervised machine learning classifications applied on the spam and ham SMS dataset.

SVM is widely used in classification and pattern recognition problems. It works well on high dimensioned data by calculating a hyper-plane that maximizes the margin between the classes causes minimize the error rate in classification problems. Its performance regarding classification is compromised when we apply it to such data, which is overlapped because this algorithm cannot maximize the margin between two classes. It is more convenient and gives better accuracy in most circumstances.

Logistic Regression (LR) is a method for examining data where one or more variables are used to produce output. Logistic regression is used to calculate the probability of class members. That is why it is considered the best learning model when there is categorical target data. It works on the relationship between independent and dependent variables.

Naive Bayes is a classification technique with a notion which defines all features is independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. Naive Bayes is a machine learning classifier which employs the Bayes Theorem.

After classification phase, the performance of individual classifiers is analyzed through performance parameters as Accuracy and Precision.

IV. RESULT ANALYSIS

The dataset used in this research is also known as Tiago's dataset. It is composed of 5,574 English, real and non-encoded messages labeled as ham (non-spam) or spam. 75% of the messages used as the training dataset and 25% was used as the testing dataset. In order to evaluate the performance of the proposed spam detection model, some metrics such as Accuracy, and Precision are used in the experiments.

Accuracy is used to calculate and measure the correctness for target classes. The highest value of this score is 1, and the lowest value is 0.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \dots .(1)$$

Precision is used to calculate and measure the correctness of classifiers. Precision can be calculated as the number of true positives divided by the sum of the number of true positives and the number of false negatives.

$$Precision = \frac{TP}{(TP+FP)} \dots .(2)$$

Where,
TP (True Positive): The proposed model predicted ham (SMS is ham), and the real value is also ham.
TN (True Negative): The Proposed model predicted spam (SMS is spam), and the real value is also spam.
FP (False Positive): The proposed model predicted spam, but the real value is ham.
FN (False Negative): The proposed model predicted ham, but the real value is spam.

Comparative performance of different Classification methods is represented in below Table 1. Fig. 2 and Fig. 3 are shows the graphical representation of Accuracy and Precision parameters respectively.

Table 1: COMPARATIVE PERFORMANCE ANALYSIS

Classifiers	Accuracy (%)	Precision (%)
SVM	97	97.2
NB	90	91
LR	89	88

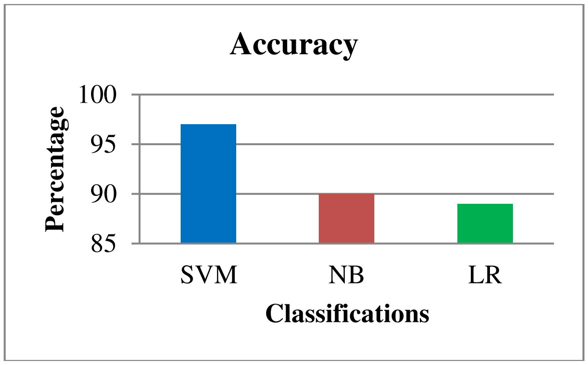

Fig. 2: ACCURACY ANALYSIS

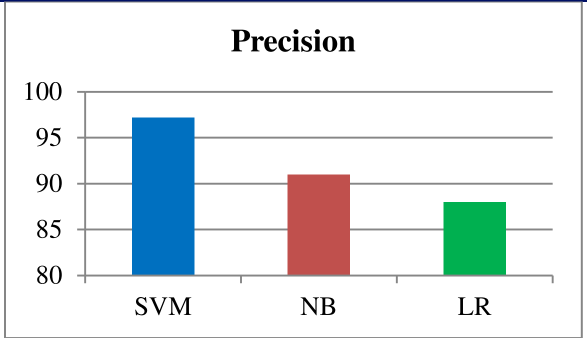

Fig. 3: PRECISION ANALYSIS

From result, it is clear that, from all supervised machine learning classifications, Support Vector Machine is efficient in terms of accuracy and precision parameters. Obtained values are accuracy as 97% and precision as 97.2%.

V. CONCLUSION

In this paper, a novel approach for Supervised Machine Learning based SMS Spam Detection is described. SMS is the most common and widely used communication network now a day. In this paper supervised machine learning classifications was applied on Tiago's dataset in order to distinguish spam from non-spam messages. Dataset is composed of mostly ham messages, hence it is described as strongly imbalanced. In order to obtain a good model, preprocessing of data was conducted. Primarily, text from the dataset is reduced to lower case, and after that tokenization was applied. 75% of the messages used as the training dataset and 25% was used as the testing dataset. The support vector Machine (SVM), Naïve Bayes (NB), and Logistics Regression (LR) are used in this study. Accuracy and Precision are used parameters in the experiments. From result, it is clear that, from all supervised machine learning classifications, Support Vector Machine is efficient in terms of accuracy and precision parameters.

VI. REFERENCES

[1] Benedict Joe Waheed Sayyed, Richa Gupta, “Social Media Impact: Generation Z and Millenial on the Cathedra of Social Media”, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Year: 2020

[2] Sandhya Mishra, Devpriya Soni, “SMS Phishing and Mitigation Approaches”, 2019 Twelfth International Conference on Contemporary Computing (IC3), Year: 2019

[3] Dimple Sharma, Aakanksha Sharaff, “Identifying Spam Patterns in SMS using Genetic Programming Approach”, 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Year: 2019

[4] Dani Gunawan, Romi Fadillah Rahmat, Arsandi Putra, Muhammad Fermi Pasha, “Filtering Spam Text Messages by Using Twitter-LDA Algorithm”, 2018 IEEE International Conference on Communication, Networks and Satellite (Comnetsat), Year: 2018

[5] A. K. Jain and B. B. Gupta, “Rule-Based Framework for Detection of Smishing Messages in Mobile Environment,” in Procedia Computer Science, vol. 125, 2018, pp. 617–623.

[6] P. Sethi, V. Bhandari and B. Kohli, "SMS spam detection and comparison of various machine learning algorithms," in International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), Gurgaon, 2017

[7] N. K. Nagwani and A. Sharaff, "SMS spam filtering and thread identification using bi-level text classification and clustering techniques," Journal of Information Science, vol. 43, no. 1, pp. 75- 87, 2017.

[8] D. Delvia Arifin, Shaufiah and M. A. Bijaksana, "Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier," in IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), Bandung, 2016

[9] T. A. Almeida, J. M. G. Hidalgo and T. P Silva, “Towards SMS Spam Filtering: Results under a New Dataset,” International Journal of Information Security Science, vol. 2, no. 1, 2013.

[10] M. Taufiq Nuruzzaman, C. Lee, M. F. A. bin Abdullah and D. Choi, “Simple SMS spam filtering on independent mobile phone,” Security and Communication Networks, vol. 5, no.10, pp. 1209–1220, 2012.