WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SMS Spam Classifier Using Machine Learning

Harsh[1], Dr. M. A. Mukunthan[2]

[1] *Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology*, Chennai, India

[2] *Faculty of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology*, Chennai, India

ABSTRACT.

In this technological age, the use of devices such as mobile phones has expanded, and the SMS service has evolved. To a multi-billion dollar industry. At the same time, the reduction in the cost of educating the administration has led to development Spontaneous Business Promotion (Spam) Shipping from mobile phones. In parts of Asia, up to 30% of instant messages were spam in 2012. Lack of genuine information base for SMS spam, text messages and restricted stimuli and their casualties Language are variables that can cause installation email filter calculations to fail to meet the expectations of their sequence. In this trust, a The database of the genuine SMS spam repository is used, and after pre-processing and highlighting, specialized ML methods and algorithms are applied. To the information base. SMS spam mail filtering is a relatively recent task to deal with such a problem.

Keywords: SMS, M, HAM, filtering, spam detection.

1. Introduction

SMS is one of the most effective forms of communication. It is based on cellular communication systems, just the cell phone needs to be in the network coverage area in order to send or receive the message. Almost everyone is using this service for communication. Various organizations deal with SMS for communicating with their clients / customers, banks and other government organizations also use SMS for communication.

Also, many business organizations use this service for advertising purposes. Thus, SMS is playing a vital role, as active internet connection is not required at all in this framework. So due to large usage of SMS, it has become one of the most favorite places for hackers and spammers. It is quiet easy for a hacker to compromise any one's cell phone just by passing or transmitting Malicious link to end user, the mobile device will automatically be compromised if end user click on the link or message being transmitted by hacker / spammer, and we can know the rest how a hacker can exploit the system if he gains control of the system.

So it has become very much important to restrict the content which the end user is receiving. So there must be a system which could tell the end user whether the received message is SPAM or not, Non SPAM message is known as HAM. So by identifying the above mentioned problems and issues, authors have developed a system which can identify whether a Message is SPAM or HAM based on the content of the message using Machine Learning technique. In this section authors have given a brief overview of Machine Learning.

Various types of Machine Learning and the techniques authors have used for developing the Machine Learning Model. Machine Learning: Machine learning is a fascinating domain as it incorporates substantial parts of different fields namely statistics, artificial intelligence theory, data analytics and numerical methods. Machine learning can be defined as semi-automated extraction of knowledge from data sets or data. Let's break down the definition into three component parts. i) Firstly, machine learning always starts with data, with an objective to extract knowledge insight from the used data or data set. ii) Secondly, machine learning involves a certain amount of automation rather than trying to gather insights from the data manually. iii) Lastly, machine learning is not fully automated i.e. it requires human interventions to make many smart decisions for the process to be successful.

Simply we can put, machine learning is an application that can improve its prediction results with successive iterations or it improves with experience. The process of an application improving with experience is, naturally enough, called Training. It can take significant iterations to gradually improve results. During the process of training, data is given to a machine-learning algorithm, which then refines its internal representation, numerical parameters, as it encounters any deviations or Training errors. The purpose of this stage is to minimize cost function, error function or maximizing likelihood by adjusting the algorithm's internal weights. When the algorithm accuracy improves, we call this learning. Once the results are accurate enough also known as scoring, the machine-learning application can be deployed to solve the problem that it was supposed to.

Machine learning is broadly categorized into two categories**:** a) Supervised Learning[3], b) Unsupervised Learning[4]. Main Categories of Machine Learning: Supervised Learning: Supervised learning also known as predictive modelling, is the process of making predictions using data. Examples of Supervised learning are Classification[5] and Regression[6]. A supervised learning Training data set is pre labelled for classification problems or function values are known in case of regression. After training is done and the model has a minimum cost function for the training data set, later switch for scoring where we can predict values for new data.

Classification: It identifies group membership. That means that if we have multiple events characterized by input parameters, which can be labelled differently, and we want our system to predict which label should be used.

Regression: Regression is a combination of multi-dimensional power supply and function interpolation. The regression problem is used to find the approximation of the function with a minimum error deviation or a cost function. In other words, the regression technique simply tries to predict numeric dependence, a function value, for example, of a data set. Figure 1. Diagrammatically shows how supervised learning is to solve problem

2. LITERATURE REVIEW

In this section, we are going to review the researches on the use of machine learning in the field of Sms Spam Classifier Using Machine Learning.

Most of the works on SMS Spam detection are content based [1], [3], [11], [12]. Content based filtering is based on the contents of SMS like spam words, unusual distribution of punctuations and message length. Yadav et al. [1] proposed a user centric approach that used content based filtering using Bayesian machine learning algorithm with user generated features like blacklisting and white listing, preferred keywords to filter unwanted SMSes and reduced the burden of notifications for a mobile user.

Narayan et al. [2] developed a two level stacked classifier to classify between spam and legitimate SMS. The first level of classifier records a subset of words whose individual probability is higher than a threshold. After that second level of classifier is invoked, this takes the chosen words form first level as input. They took different combinations of machine learning classification algorithms in two levels such as Bayesian and SVM, SVM and Bayesian, Bayesian and Bayesian, SVM and SVM.

Ishtiaq et al. [3] proposed a SMS spam classification algorithm using the combination of Naive Bayes classifier and Apriori algorithm. They integrated association rule mining using Apriori algorithm with Bayesian algorithm. Apriori retrieves the most frequent words occurred together then Bayesian calculates the probability of occurring a word independently and together with other words, in spam or ham messages.

Gomez et al. [4] analysed to what extent Bayesian filtering techniques used to block email spam, can be applied to the problem of detecting and stopping mobile spam. They pre-processed the messages with different tokenization approach, selected features and tested them with different machine learning algorithms, in terms of effectiveness. They demonstrated that Bayesian filtering techniques can be effectively transferred from email to SMS spam with appropriate feature extraction.

Many proposed techniques used non-content based filtering [5], [7]. Warade et al. [2] detected the spam messages by checking mutual relation between the sender and receiver and the content of the messages. If no mutual relation is found between sender and receiver and message contains spam contents, then the system tags the message as spam and sends it to spam box. If mutual relation and no spamming content exist then it directly sends to inbox of the receivers mobile. It solved the problem of balance deduction and wastage of SMS memory. But calculating only mutual relation is not a proper solution. Spam detection algorithm needs both classification algorithm and this kind of feature extraction from contents.

Qian Xu et al. [6] investigated ways to detect spam message senders based on non-content features that include temporal and graph-topology information but exclude contents because of user-privacy issues. They focused on the problem of identifying professional spammers based on the overall message sending patterns. Furthermore, they only concentrated on finding SMS spam on the server side, as the client-side detection is mostly content based.

Sharaff et al.34 proposed a [7] novel SMS spam filter model based on a biologically inspired algorithm named krill herd optimization and dendritic cell algorithm. Their experimental results showed that the proposed model gave more accurate results compared with other ML classifiers like NB, LR, SVM, and XgBoost classifier. Another study from Bosaeed et al.35 developed a multi-filter that applied multiple ML based classifiers using three classification methods, namely Naïve Bayes (NB), SVM, and Naïve Bayes Multinomial (NBM). The study shows the flexibility of multiple platforms by implementing their proposed model partly and fully on both mobile and server apps, thus ensuring computational resource optimization.

Alzahrani and Rawat 36 also [8] presented a comparative study of different ML algorithms for SMS spam detection. Four ML algorithms were applied, and the best performance was achieved with the neural network algorithm compared with other classifiers. A similar study was carried out by authors Theodorus et al. 37 that compared the performance of eight ML classifiers in Bahasa Indonesia SMS text classification. Other applications of ML algorithms have been presented by different works of literature like the Naïve Bayes algorithm, 27,38–40 neural network classifier, 41 self organizing map,15 KNN, H20 framework, 42 and so on. Ensemble learning was applied by Sisodia et al. The authors presented an automated framewoark for SMS spam classification using various classifiers, namely individual classifiers such as KNN, NB, SVM, ID3, CART, C4.5, and ensemble classifiers such as Adaboost, random forest, and voting. The experimental results showed promising results with the best accuracy obtained with the ensemble learning classifiers based on random forest.

Sharma and Sharaff 43 [9] recently considered a different perspective, which applied genetic programming to SMS spam filters to reduce false-positive errors. The performance of the proposed method shows significant improvement in the classification of SMS spam with the increas-ing number of generations. Other evolutionary methods also gained effective relevance in the SMS spam classification; studies from Al-Hasan and El-Alfy2 proposed a novel approach DCA on NB and SVM algorithms using various feature sets. Onashoga et al. 44 developed a collaborative and adaptive filtering system based on an artificial immune system similar to Mahmoud and Mahfouz, 45 which also applied artificial immune system for SMS spam.

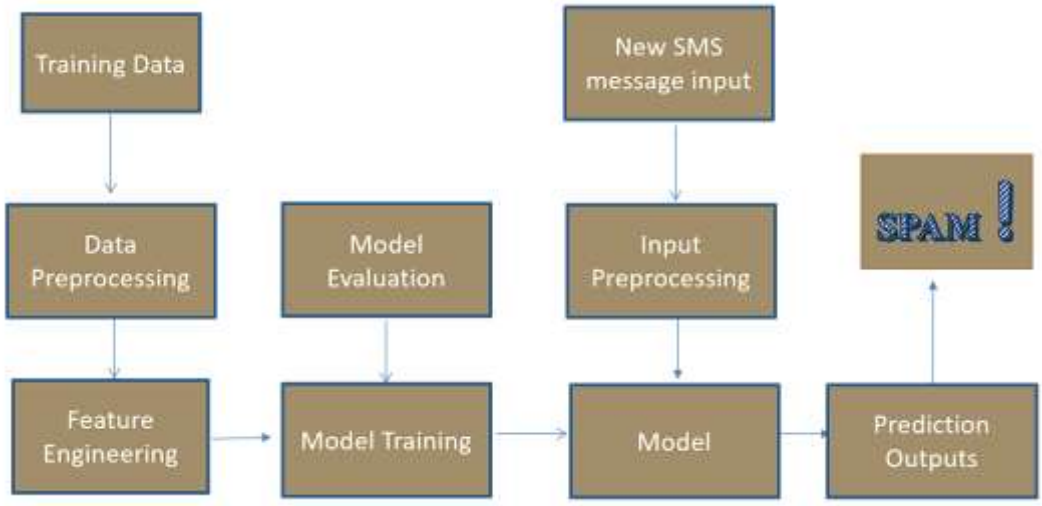
Authors presented the application of deep learning techniques for SMS spam detection using [10] LSTM in Gadde et al. 46 and Al-Bataineh and Kaur. 47 Authors in the former also applied three different word embedding techniques based on the count, TF-IDF, and hashing vectorizer. The experimental results for LSTM were compared with some of the state-of-the-art ML techniques. However, authors in the latter demonstrated the robustness of LSTM topologies with a clonal selection algorithm for text classification. The study was evaluated using three datasets and benchmarked with some of the state-of-the-art ML classifiers. The experimental results showed that their proposed model outperforms other models regarding accuracy, precision, recall, F1-score, and computational time. Roy et al. 48 also proposed a deep learning method based on convolutional neural networks and long short-term memory models in the classification of SMS spam. The study showed a significant performance of deep learning models using three configurations, and the addition of regularization parameters such as dropout improved classification accuracy. Another interesting work by Xia and Chen49 introduced an enhanced Hidden Markov Model for a weighted feature and label words. Their study shows that the application of weighted features enhanced HMM outperforms the LSTM with respect to accuracy and computational speed.

A new method based on a lightweight deep neural model referred to as Lightweight [11] Gated Recurrent Unit (LGRU) was presented by Wei and Nguyen50 for SMS spam detection. In order to show the effectiveness of the proposed method, the authors compared their results with over 30 different machine and deep learning classifiers. Aside from that, the proposed method achieved better results compared with existing models; the authors also claimed that it also incurs less complexity in terms of training time. Authors Annareddy and Tammina 51 and Huang 52 applied deep learning models for the detection of SMS spam. The former showed a comparative study of two deep learning models based on a convolutional neural network and RNN on a large SMS corpus. The overall results show interesting findings. At the same time, the latter applied the CNN model on the Chinese Wikipedia corpus. The authors further demonstrated the impact of hyper-parameters in achieving optimal results. Another interesting finding that combined two deep learning models, namely RNNs and the LSTM model, was proposed by Chandra and Khatri. 53 The study showed that the proposed model could predict based on previous knowledge of patterns and the current vector set. The study concluded with a significant improvement in performance based on accuracy with excellent and acceptable runtime.

Lee and Kang 54 [12] introduced word embedding methods for building a feature vector and applied deep learning methods for SMS spam classification. The experimental result shows little improvement in the performance of the deep learning method compared with the conventional ML method of SVM-light. Chen et al.3 and Baaqeel and Zagrouba 55 introduced a hybrid system for SMS spam classification. The former applied a trust manage-ment scheme using behavioral, and SMS traffic data and the article concluded that the proposed prototype achieved effective detection with respect to efficiency, robustness, and accuracy. However, the latter combined six supervised models and unsupervised methods for SMS spam detection. However, the overall experiments showed that the combination of K-means with SVM gave an outstanding performance based on accuracy. Thus, the study concluded that the hybrid system performed way better than the single classifier.

3. METHODOLOGY

A. Architecture Diagram



B. Dataset

We will classify spam and ham messages using the Keggle SPAM/HAM Dataset, which consists of total of 5573 messages with different input for spam and ham messages. It also contains 2 different columns v1 for ham/spam and v2 for messages. The dataset has been prepared by extracting features from each image into csv file.

C. Data Visualization

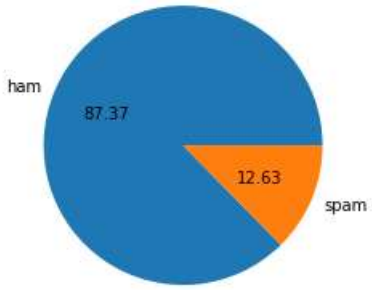


Fig 1: Percentage of Ham against spam data in dataset

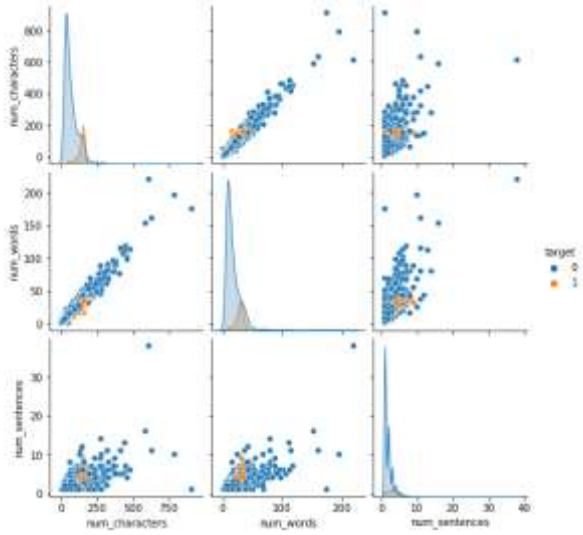


Fig 2: Ham and Spam data distributed in two classes

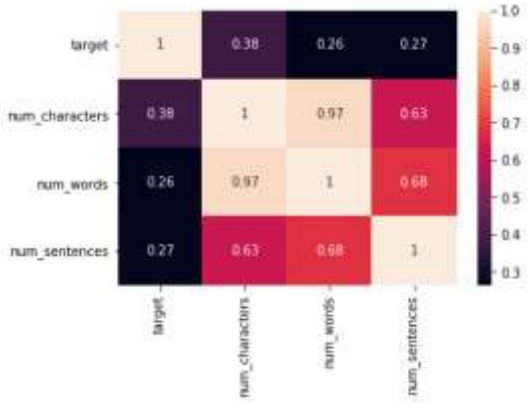


Fig 3: correlation of each feature with the target

D. Algorithms

Multinomial Naive Bayes is a useful algorithm in many respects, especially for solving low-data text classification problems. The way in which it can make accurate predictions with limited training data by assuming naïve independence of words is its main advantage, but ironically also a disadvantage when comparing it to more advanced forms of NLP. Other approaches, such as classifiers which use word embeddings, can encode meaning such as context from sentences to obtain more accurate predictions. Still, there is no denying that Naïve Bayes can act as a powerful spam filter. With an ever-growing amount of textual information stored in electronic form such as legal documents, policies, company strategies, etc., automatic text classification is becoming increasingly important. This requires a supervised learning technique that classifies every new document by assigning one or more class labels from a fixed or predefined class. It uses the bag of words approach, where the individual words in the document constitute its features, and the

order of the words is ignored. This technique is different from the way we communicate with each other. It treats the language like it's just a bag full of words and each message is a random handful of them. Large documents have a lot of words that are generally characterized by very high dimensionality feature space with thousands of features. Hence, the learning algorithm requires to tackle high dimensional problems, both in terms of classification performance and computational speed.

Decision Tree algorithm mostly used for regression and classification problems. It is a tree-based model that is used to make decisions based on a set of input features. In the case of spam and ham detection, decision predict the based on the features from mammogram images. In decision tree, the input features are split into nodes based on their importance in predicting the outcome variable. The decision tree algorithm uses a top-down approach to split the features into nodes, with each node representing a decision rule. The algorithm continues to split the nodes until it reaches a point where further splitting does not affect. The decision tree algorithm will be trained on training set using the features of mammogram images. The objective is to create a tree that can accurately classify s into benign or malignant categories. The decision tree algorithm selects the most important features are then used to create a tree structure that consists of decision nodes and leaf nodes. The decision nodes are based on the feature values and determine the next node in the tree, while the leaf nodes represent the final decision or prediction. Once the decision tree is created, it is evaluated on the test set to determine its accuracy in classifying s and avoid false positives and false negatives. Decision trees have several advantages for Spam and Ham detection. They are easy to understand and interpret, making it possible for professionals to interpret the results and make informed decisions. These trees can also handle numerical and categorical data, making them versatile for a wide range of applications. However, they are prone to overfitting and may not perform well on datasets with noisy or irrelevant features.

Random Forest, usage of this as a machine learning algorithm is prevalent in various tasks such as classification, regression, and others. It is a type of ensemble learning method that uses multiple decision trees to make predictions. The algorithm works by creating a large number of decision trees and combining their predictions to make a final prediction. Each tree is built using a subset of data and features, which helps to reduce overfitting and increase accuracy. It is often used in combination with other algorithms to improve performance even further, such as in a stacked ensemble approach. The algorithm works by analyzing a set of input variables, such as the size and shape of the, and using these variables to construct a decision tree. The decision tree algorithm selects the most important features are then used to create a tree structure that consists of decision nodes and leaf nodes, The stopping criterion is typically based on the purity of the subsets, or the degree to which all members of a subset belong to the same class. Random Forest algorithm can be attributed to its ability to handle high-dimensional datasets and reduce overfitting. It also has the advantage of combining multiple decision trees to improve the prediction accuracy. This is typically done by taking the majority vote of the predictions, although other methods can also be used. The result is a highly accurate prediction of is benign or malignant. One of the advantages of Random Forest is its ability to handle large datasets with many input variables. The algorithm can handle both categorical and continuous variables, and can also handle missing data. It is also robust to outliers and noise in the data, and can handle nonlinear relationships between the input variables and the output variable. In terms of performance, Random Forest has been shown to outperform many other machine learning algorithms on a variety of datasets, including the Spam and Ham Wisconsin dataset. It is often used in combination with other algorithms to improve performance even further, such as in a stacked ensemble approach.

	Algorithm	Accuracy	Precision
1	KN	0.900387	1.000000
2	NB	0.959381	1.000000
8	ETC	0.977756	0.991453
5	RF	0.970019	0.990826
0	SVC	0.972921	0.974138
6	AdaBoost	0.962282	0.954128
10	xgb	0.971954	0.950413
4	LR	0.951644	0.940000
9	GBDT	0.951644	0.931373
7	BgC	0.957447	0.861538
3	DT	0.935203	0.838095

Fig 4: classification report of all algorithms

CONCLUSION

From the above discussion and experimentation have concluded that Machine Learning algorithms can play a vital role in identifying SPAM SMS. The accuracy obtained in this work is more than 95% in all the cases.

Random Forest and Decision Tree both performed well with accuracies of 97.1% and 93.7% respectively. However, MultinomialNB was better the other two algorithms with an accuracy of 95.5%. This indicates that MultinomialNB is better suited for spam and ham detection tasks and can provide more

accurate results. The high accuracy and precision of the MultinomialNB algorithm can be attributed to its ability to handle high-dimensional datasets and reduce overfitting. It also has the advantage of combining multiple decision trees to improve the prediction accuracy.

This makes MultinomialNB a powerful tool for spam and ham detection. In conclusion, this paper highlights the importance of machine learning in diagnosis and provides insights into the performance of different algorithms for spam and ham detection. The results suggest that MultinomialNB can be an effective tool for professionals in diagnosing spam and ham and can potentially lead to earlier detection and improved outcomes.

ACKNOWLEDGMENT

We would like to thank Dr. M. A. Mukunthan for his expertise and guidance as a Faculty of Computer Science and Engineering at Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology in Chennai, India have been instrumental in shaping the direction and quality of our publication.

References

Njoku, Mary Gloria. (2015). The use of short message service in post-secondary education.

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, “Foundations of Machine Learning”, The MIT Press ISBN 9780262018258, 2012.

Sumit Das, Aritra Dey, Akash Pal, Nabamita Roy, “Application of Artificial Intelligence in Machine Learning: Review And Prospect”, International Journal of Computer Applications”, Volume 115, Number 9. 2015.

S.B. Kotsiantis. “Supervised Machine Learning: A Review of Classification Techniques” Informatica 31 (2007) 249-268

Nanhay Singh, Ram Shringar Raw, Chauhan R.K. ,” Data Mining With Regression Technique”, Journal of Information Systems and Communication ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, 2012, pp.-199-202

Amandeep Kaur Mann & Navneet Kaur, “Review Papers on Clustering Techniques”, Global Journal of Computer Science and Technology Software & Data Engineering, Volume 13, Issue 5, Version 1.0, 2013.

Ashima Sethi, Prerna Mahajan, “Association Rule Mining: A Review”, The International Journal of Computer Science and Application, Volume 1, No. 9, November 2013.

Swati Gupta, “A Regression Modeling Technique on Data Mining”, International Journal of Computer Applications, Volume 116, No. 9, April 2015.

Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

M. Nivaashini, R.S.Soundariya, A.Kodieswari, P.Thangaraj, SMS Spam Detection using Deep Neural Network, International Journal of Pure and Applied Mathematics, Volume 119 No. 18 2018, 2425-2436 ISSN: 1314-3395 (on-line version), url: http://www.acadpubl.eu/hub/, Special Issue

Dipak, R & Kawade, Dipak & Oza, Kavita. (2018). CONTENT-BASED SMS SPAM FILTERING USING MACHINE LEARNING TECHNIQUE. 12.

P. Navaney, G. Dubey and A. Rana, "SMS Spam Filtering Using Supervised Machine Learning Algorithms," 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, 2018, pp. 43-48. doi: 10.1109/CONFLUENCE.2018.8442564.

Behera, Bichitrananda & Kumaravelan, G.. (2019). Towards the Deployment of Machine Learning Solutions for Document Classification. International Journal of Computer Sciences and Engineering. 7. 193-201. 10.26438/ijcse/v7i3.193201.

Bichitrananda Behera, G.Kumaravelan. (2020). Performance evaluation of Machine learning algorithms in Biomedical Document Classification. International Journal of Advanced Science and Technology,29(05),5704-5716