ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Comparative Study of Deep Learning, Random Forest and Naive Bayes Approaches in SMS Spam Detection

Bhavya K[1], Rafeeque P C[2]

PG Student, Dept of Computer Science & Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Head of the Department, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: Any unwanted or useless text message delivered to a mobile phone through Short Message Service (SMS) is called as a spam SMS. The spam SMS issue is gradually increasing along with the increase in the use of text messaging. Users usually do not like receiving such messages as they are just disturbing, and here arises the need for spam filters. The proposed system focuses on detecting the spam messages by identifying the features of each messages contained in UCI machine learning repository. A message contains different valid features, making it as either spam or ham. The proposed method makes use of three machine learning approaches called Deep Learning, Naive Bayes, and Random Forest approach. Finally, a comparison is made among the three approaches in order to identify which technique gives the best performance in this particular task. Interestingly, each of them are so close to each other based on their performance, and gives a promising result in the task of SMS spam detection.

Keywords: SMS Spam Detection, Natural Language Pro- cessing (NLP), Deep Learning, Naive Bayes, Random Forest

I. INTRODUCTION

Short Message Service (SMS) could gain the attention of people all over the world in parallel with the increased popularity of mobile phone devices. The main reason is the reduction in the cost of messaging services, and thus it resulted in growth in unsolicited commercial advertisements (spam) being sent to mobile phones. Many email filtering algorithms existing today under perform in the classification of spam messages because of the lack of real databases for SMS spam, short length of messages and limited features, and their informal and unstructured language. There exist major differences between spam-filtering in text messages and emails. In case of emails, there exist a variety of large data sets, on the other hand, real databases for SMS spam are very limited. Due to the small length of text messages, the number of features that can be used for their classification is much smaller than that of emails. Additionally, no header exists as well. One more reason is that text messages are full of abbreviations and language used by people would be informal in most of the cases, which we can less see in emails. All of these factors may result in serious flop in performance of major email spam filtering algorithms in the task of classification of short text messages into spam and ham. The importance of SMS spam detection occurs in many aspects. One of the main goal of spam filters is to identify and block the spam messages as it is a hectic disturbance to users. The spam messages might be in the form of notifying users about any offers sent from the mobile companies to which user's number is connected with, might be from any banks in which the particular user has account, or it can be any kind of advertisements too. Some of the spam messages might harm the device when it contain malicious links and all, which can even steal the personal information, identity or some valid credentials of a user. Thus it is very much essential to have a spam filter.

There are various techniques used for SMS spam detection, such as using Support Vector Machine (SVM) [1], k Nearest Neighbour (KNN) [2], Naive Bayes (NB) [3], artificial neural network [4], decision tree [5] and random forest [6]. Different comparisons and experiments were made between different techniques using different data sets, and their results could show that SVM and NB classifiers provided highest accuracy, also classifiers that use decision tree, Bayesian classification and logistic regression still suffer from increased running time. The proposed framework, focuses on detecting spam messages from the entire dataset, which is the UCI machine learning repository [7], based on the features of messages. As we have noticed, the spam messages would contain certain different features like existence

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

of URLs, phone numbers, special characters etc. Also the number of characters, words etc contained in the message would also contribute a lot for a message to be ham or spam.

In the proposed system, a spam detection system is introduced which is built by three classifiers that make use of techniques like Deep Learning (DL) [8] (Sequential Model), Random Forest (RF) and Naive Bayes (NB). Moreover, the matrices that will be used for evaluating the model are the accuracy, precision, recall, f-measure in addition to measuring the time efficiency. In addition, the dataset used for experiments is the same dataset that is available in UCI Machine Learning Repositories. Consequently, the dataset will be explored and Python language will be used to make pre-processing. After considering the performance matrices of each classifier, the best algorithm is chosen.

The rest of the paper is organized as follows: Section II covers the works which are related to the SMS spam detection task. Section III discusses the methodology adopted for implementing the proposed system. Section IV is all about the results obtained for the proposed methodology. Section V concludes the discussion , and finally added various references used for the study.

II. RELATED WORKS

Tiago A. Almeida et. al. [9] proposed an approach in which a real, public and non-encoded SMS spam collection is used. Two tokenizers were introduced here, in The first tokenizer considers tokens start with a printable character followed by any number of alphanumeric characters, excluding dots, commas and colons from the middle of the pattern. With this pattern, domain names and mail addresses would be split at dots, so that the classifier can recognize a domain even if sub-domains vary. Coming to the second tokenizer, any sequence of characters separated by blanks, tabs, returns, dots, commas, colons and dashes are considered as tokens and this simple tokenizer intends to preserve other symbols those may help to separate spam and legitimate messages.

The system proposed by Qian Xu et. al. [10] uses a real- world data set from a large telecommunications operator in China, and examines the effectiveness of various content- less features that range from network and to time-oriented categories. This approach points out that some intuitively appealing features are in fact not much effective, instead a combination of temporal and network features can be much more useful in training high performance classifiers for spammer detection.

Dr.Ghulam Mujtaba et. al. [11] proposed a work which describes a mobile station based approach where a spam sms would be identified and removed as soon as it is received at the mobile device. Four features were extracted here. These features are existence of frequently occurring diagrams in the message and message class, also the size of the message and existence of frequently occurring monograms in the message. The performance of Naive Bayes algorithm is shown to be the better one when compared to other algorithms explored. The other algorithms used were Artificial Neural Networks and Decision Tree classifier.

The approach put forward by Wei Li and Sisheng Zeng used Vector Space model based on Spam SMS filtering. It addressed particularly of Short message Service ,such as short, vocal , domain related etc. It used much modification on the traditional SVM model. This technology has been deployed in Production environment of Dahan Tricom Corparation and results in Production Department turn out be Applied in SMS Commercial Companies.

Agarwal et. al. [12] proposed an approach which mainly focused on spam detection for Indian messages. It was purely a content based method where different machine learning algorithms such as Multinomial Naive Bayes(MNB), Support Vector Machine(SVM), Random Forest(RF) and Adaboost were used and compared.

El-Alfy and AlHasan [13] have proposed a model for filtering text messages for both email and SMS. They analyzed different methods in order to finalize a feature set in order to reduce the complexity. They used two classification algorithms; Support Vector Machine (SVM) and Naive Bayes, which are trained using feature vector made up of 11 features like URLs, likely spam words, emotion symbols, special characters, gappy words, message metadata, JavaScript code, function words, recipient address, subject field and spam domain. Also, they evaluated their proposed model on five email and SMS datasets.

Jialin et. al. [14] proposed a message topic model (MTM) for filtering Spam messages. MTM or Messages Topic Model considers symbol terms, background terms and topic terms to represent spam messages which are based on the

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
PCON-2019
National Conference and Seminar on Innovations in Engineering & Technology
Govt Polytechnic College, Palakkad
Vol. 8, Special Issue 1, January 2019

probability guess of latent semantic analysis. In order to remove the sparse problem, they used k-means algorithm by training SMS spam messages into random irregular classes and then aggregating all SMS spam messages as a single file to capture word co- occurrence patterns.

Neelam Choudhary and Ankit Kumar Jain [15] proposed an approach that can detect and filter the spam messages using machine learning classification algorithms. The characteristics of spam messages were studied in depth and then found ten features from them, which was then used to train the model and thereby efficiently filter SMS spam messages from ham messages.

Chen et. al. [16] proposed a PBS(Pseudo base station) detecting and tracking system is designed and implemented, by conducting topic analysis of messages received by cell- phones and analysing their temporal and spatial distribution patterns. Using the system, a variety of exploratory analysis, including categorizing PBSes into either stationary or moving PBSes were performed, also discovering and visualizing their behaviour patterns, and identifying districts that tend to suffer from a particular type of fraud messages were also done.

Dima Suleiman and Ghazi Al-Naymat [17] extended the previous works done on sms spam detection by extracting the maximum possible features from the messages included in the selected dataset. The classifier proposed was depending on H2O framework, where strong machine learning algorithms are used in order to improve the performance of the system, and best approach among them was identified. They could found random forest as the best algorithm compared to other algorithms used.

III. METHODOLOGY

The proposed spam detection system is based on machine learning approaches. The architecture and the important steps in building the proposed system is described here.

A. Architecture

Fig 3.1 shows an overall work flow or architecture of the proposed spam detection system for short message services. The main processes in the proposed system are dataset collection, data pre-processing, feature extraction and model training. The dataset chosen is UCI machine learning repository. The first task is to pre-process the selected dataset. The pre- processed dataset is then used in the feature extraction phase (in case of NB and RF) , where various features are extracted from it. The extracted features are used to train the system. The trained system will be able to predict whether a particular input is spam or ham. In case of DL, there doesn't occur the need of feature extraction as the features would be automatically selected by the system.

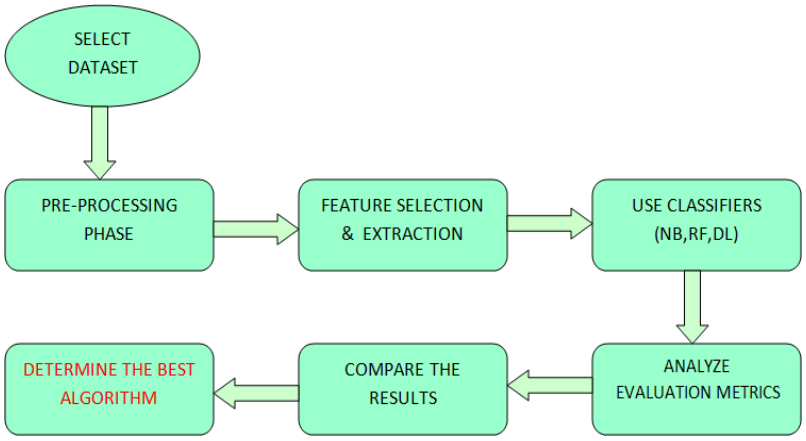

Fig 3.1: Architecture of SMS Spam Detection System

B. Dataset Collection

In the future, timeline for a non-human actor can be generated, e.g., a timeline of art or science in the Renaissance. Also aim to define annotation guidelines. Also aim to define annotation guidelines for annotation of historical events and release a much larger annotated dataset that can be used for various tasks such as entity/event extraction and segmentation, co-reference resolution of named entities as well as events.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Table I shows an example of ham and spam messages in the chosen UCI dataset.

Table I : Data From UCI Machine Learning Repository

Ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
Ham	Ok lar... Joking wif u oni...
Spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply

C. Data Pre-Processing

When the task reaches to data pre-processing, the only thing to be done is removal of stop words, as they cannot be considered as features. As the detection of spam messages purely depends on the features of messages, the important constituents of a sentence cannot be removed in pre-processing phase. Thus we are not removing any punctuation marks or special characters and all. Stemming, and mapping of short words to their original words are not at all essential as they would not help us a lot to detect spam, and the result would be close enough even though these steps are not done.

D. Feature Extraction

The pre-processed dataset is then given to the feature extraction phase where the major features of messages that can be used for classifying spam and ham messages are chosen or extracted. The major features extracted from messages are as listed and explained below.

1) Presence of Mathematical Symbols : Most of the spam messages usually contain mathematical symbols as per the scenario. For example, the symbol + can be used for free services messages. Mathematical symbols considered in our experiment are +, , < , > , / and .

2) Presence of URLs : We consider the presence of URLs as a major feature since harmful spam SMS contains URLs. The user would be asked to visit the URLs and once they click on the link, it may capture users' personal information, debit/credit card details, any passwords etc in an unhealthy manner and sometimes ite may cause downloading of some file (may contain virus).

3) Presence of Dots : The presence of dots usually indicates that it is a legitimate message, because while chatting, people often use dots.

4) Presence of Special Symbols : Spammers use special symbols in messages for various reasons. For example, in fake award messages, in order to represent money in the dollar, special symbol $ is being used. Similarly the symbol ! is used for seeking the special attention of users with the usage as in CONGRATULATIONS! WINNER!, etc. Special symbols that are considered in this approach are !, , , $ etc.

5) Presence of Lower-cased Words : Checks for all the lower-cased words their presence in a message can be used to seek users attention.

6) Presence of Upper-cased Words : The presence of upper- cased words are considered as a major feature as spammers usually use upper-cased words to seek users attention. For example, RINGTONE, ATTENTION, WON, PRIZE, FREE etc.

7) Presence of Mobile Numbers : The presence of mobile number in a message is considered as a feature in order to identify spam messages, because spammers usually give mobile number in a message. They ask the users to dial on the given number, attacker on the other side ask for users personal details, bank details, etc. Let us consider an example: you have won a 2,050Rs price! To claim, call 09050080301.

8) Presence of Specific Keywords : Presence of some specific keywords like awards, won, send, ringtone, free, service, lottery, video, visit, congrats, Please, delivery, cash, claim, Prize, delivery, etc. are considered as spam keywords because they are usually used to attract users.

9) The Message Length : It includes the total length of the message including smileys, space, symbols, special characters etc. 160 characters is the text limit of SMS messages.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Table II: Sms Message Feature Value for Ham and Spam Messages

Feature Type	Have you finished your work ? (Ham message)	CONGRATULATIONS!! You have won Nokia 3650. Call 09066382422 to claim your prize. It's your final chance!! (Spam message)
Presence of mathematical symbols	0	1
Presence of URLs	0	0
Presence of dots	0	0
Presence of special symbols	0	0
Presence of Lower-cased Words	1	1
Presence of Upper-cased words	0	1
Presence of mobile number	0	1
Keyword specific	0	1
Message length	30	141

E. Machine Learning Algorithms

After the features are extracted, the next step is to apply machine learning algorithms and train the classifiers. Once the classifiers are trained, test with a new instance. Naive Bayes, Random forest and Deep Learning are the machine learning algorithms used here.

F. Evaluation metrics

The proposed classification system can be evaluated using five metrics which are as follows: Precision, Recall, F-measure and Accuracy. As it is a binary classification problem, confusion matrix can be used. For the computation of the metrics, four identifiers are to be defined, and they are: true positive, false positive, true negative, and false negative values. Accuracy is the percentage of correctly classified messages out of total number of messages. The number of classified messages that are actually spam gives us the value of precision, recall refers to the number of the spam messages that are correctly classified as spam itself. Now, f-measure is the one which combines precision and recall into one measure. Accuracy and f-measure values are supposed to be high in order to get a better classification result.

IV. RESULTS AND DISCUSSIONS

Various experiments are performed to evaluate the performance of the proposed SMS Spam detection system. Initially features are selected on the basis of behaviour of spam and ham messages and then extracted these features from the dataset to get the feature vector. After extracting features from the dataset, various classification algorithms such as Naive Bayes, Random Forest are being applied to get the performance metrics.

Table III : Performance Results of Naive Bayes & Random Forest Classifiers

Classifier	**Precision**	**Recall**	**Accuracy**	**Average Precision**	**F-Measure**
Naive Bayes	79.87	83.43	94.70	82.82	81.61
Random- Forest	95.12	85.53	97.68	86.66	90.26

Table III shows the comparison between NB and RF classifiers based on their performance in the proposed approach. When compared to the performance of Naive Bayes classifier, Random Forest gives best performance based on the selected features. All the measures are given in the percentage form.

Coming to the deep learning approach, Keras' [18] Sequential() [19] is used, which is a simple type of neural network that consists of a stack of layers executed in order. Even a stack of only two layers (input and output) can be used to make a complete neural net, but it cannot be considered as a deep neural network. Here we're inputting a sentence which is then converted to a one-hot matrix of defined length (here it is 3000). The parameters like, how many outputs we want to come out of that layer and what kind of activation function to use are also included. Activation functions usually differ, mostly in speed, but all the ones available in Keras and Tensor Flow [20] are feasible, and the one that is used in the first layer is relu. The out network mostly consists of dense layers, the standard, linear neural net layer of inputs, weights, and outputs. Also, the output layer consists of two possible outputs, spam and ham. In between the input

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

and output layers, one more dense layer and two dropout layers are used. The data is evaluated in groups of batches, sizing 32 instead of the size of entire dataset. This could make the networks get trained much more quickly. epochs is how many times we need to do this batch-by-batch splitting. Initially, it is set to 7, but caused overfitting. 5 is found to be good in this case. Verbose is set to 1, so that it will display an animated progress bar that indicates the training progress of each epoch. Also, the optimizer used here is Adam Optimizer Finally, 10 % (0.1) of the training data is used for the validation. Table IV shows the major aspects considered in the deep learning approach.

Table IV : Aspects used in Deep Learning Approach

Model	**Keras' Sequential**
Activation Functions	relu, softmax, sigmoid
Optimizer	adam
Batch Size	32
Verbose	1
Epochs	5
Validation Split	0.1

Considering all these aspects, the approach gives the best result among the three approaches used for the task is sequential neural network model. The accuracy obtained is 98.83%, which is the best compared to the other two. The performance of proposed system is compared with the baseline system and is given in Table V.

Table V : Comparison with Baseline System

Model	**Best Approach Chosen**	**Accuracy**
Baseline System	Random Forest	97.7%
Proposed System	Neural Network Model	98.83%

In the baseline system [17], the experiment was done using H2O framework, and the best algorithm chosen in terms of precision, recall, f-measure and accuracy was the Random Forest which could achieve better results with values equal to 96%, 86%, 91% and 97.7% respectively. On the other hand, the proposed system gives the best performance compared to the baseline system, when it is modelled using neural network model with an accuracy of 98.83%.

CONCLUSION AND FUTURE WORKS

The proposed system is able to detect spam messages. Two of the selected machine learning approach use some excellent features which help to make the spam messages stand out from the entire dataset. On the other hand, by using deep learning in classification, the features are automatically selected for detecting spam messages. However, the experiment shows that even if deep learning technique takes more run time, it is able to give the best result. Random Forest is ranked second, and that of Naive Bayes is third, based on the evaluation metrics. SMS Spam detection is becoming more essential in the modern era, and the proposed system offers best performance in the same task compared to the previous systems. The selected features contribute a lot in distinguishing a spam and a ham message. Sometimes, the features which best contribute to a spam might also be the features of some ham message and vice verse. Thus, more strong features can also be found out, and can develop even more powerful system.

REFERENCES

[1]. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Sup- port vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.
[2]. L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
[3]. K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, vol. 18, 2006.
[4]. R. J. Schalkoff, *Artificial neural networks*, vol. 1. McGraw-Hill New York, 1997.
[5]. J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.
[6]. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[7]. A. Asuncion and D. Newman, "Uci machine learning repository," 2007.
[8]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

[9]. T. A. Almeida, J. Almeida, and A. Yamakami, "Spam filtering: how the dimensionality reduction affects the accuracy of Naive bayes classifiers," *Journal of Internet Services and Applications*, vol. 1, no. 3, pp. 183–200, 2011.
[10]. Q. Xu, E. W. Xiang, Q. Yang, J. Du, and J. Zhong, "Sms spam detection using noncontent features," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 44–51, 2012.
[11]. G. Mujtaba and M. Yasin, "Sms spam detection using simple message content features," *J. Basic Appl. Sci. Res*, vol. 4, no. 4, pp. 275–279, 2014.
[12]. S. Agarwal, S. Kaur, and S. Garhwal, "Sms spam detection for indian messages," in *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on*, pp. 634–638, IEEE, 2015.
[13]. E.-S. M. El-Alfy and A. A. AlHasan, "Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm," *Future Generation Computer Systems*, vol. 64, pp. 98–107, 2016.
[14]. J. Ma, Y. Zhang, J. Liu, K. Yu, and X. Wang, "Intelligent sms spam filtering using topic model," in *Intelligent Networking and Collaborative Systems (INCoS), 2016 International Conference on*, pp. 380–383, IEEE, 2016.
[15]. N. Choudhary and A. K. Jain, "Towards filtering of sms spam messages using machine learning based technique," in *Advanced Informatics for Computing Research*, pp. 18–30, Springer, 2017.
[16]. Z. Chen, "Malicious base station and detecting malicious base station signal," *China Communications*, vol. 11, no. 8, pp. 59–64, 2014.
[17]. D. Suleiman and G. Al-Naymat, "Sms spam detection using h2o framework," *Procedia Computer Science*, vol. 113, pp. 154–161, 2017.
[18]. F. Chollet *et al.*, "Keras," 2015.
[19]. H. Wang and D. Bell, "Sequential neural network model," in *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pp. 22–26, IEEE, 1994.
[20]. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin,
[21]. S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning.," in *OSDI*, vol. 16, pp. 265–283, 2016.
[22]. N. Chaudhari and V. Jayvala, "Survey on spam sms filtering using data mining techniques," *International Journal of Advanced Research in Computer and Communication Engineering ISO*, vol. 3297, 2007.