A Review on Various Approaches on Spam Detection of Mobile Phone SMS

Samadhan M. Nagare, Research Student, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, samadhannagre340@gmail.com

Pratibha P. Dapke, Research Student, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, pratibhadapke189@gmail.com

Syed Ahteshamuddin Quadri, Research Student, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, syedahtesham1432@gmail.com

Sagar B. Bandal, Research Student, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, sagarbandal2901@gmail.com

Manasi Ram Baheti, Assistant Professor, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, mrb.csit@bamu.ac.in

Abstract-Spam and ham SMS detection in mobile phones, this paper presents review of spam Due to the availability of inexpensive bulk SMS bundles and the fact that messages elicit greater response rates due to the one-on-one and personalized nature of the service, they are a modern problem. In this study, to differentiate the messages, we will that will be classified into two categories spam and ham. Dataset of messages that contain whether or not the records are authentic messages is indicated by the text of SMS messages at the side of the label. Spam is defined as a dataset that includes SMS message text content and a label designating it as junk mail. In SMS spam messages, marketers use SMS text messages to send unwanted advertisements to specific clients. To get around this, we use the SMS spam dataset to compare the machine learning methods used to detect spam and non-spam messages and to determine the accuracy threshold.

Keywords: SMS Spam Ham, SMS Detection

I. INTRODUCTION

The most popular and commonly utilized kind of communication is short message service. In different regions of the world, the term "SMS" is used to refer to both user activity and all forms of brief text messaging. It is being used as a tool for online offers, financial updates, agricultural information, and product advertising and promotion. SMS marketing, often known as direct marketing, uses SMS technology. occasionally, SMS advertising is an issue that agitates users. is an issue that agitates users. This SMS are referred to as spam SMS. One or more unsolicited messages that are sent or posted as part of a big collection of messages with essentially the same content is referred to as spam. SMS spam is sent with the intention of disseminating inappropriate pornographic content, internet offers, political difficulties, and advertisements for various products. Because of this, spam SMS flooding has grown to be a significant issue. all throughout the world. Owing to SMS communication's rising popularity,

Compared to other forms of spam, such as Twitter, email, and mobile phone SMS, SMS spam has become more common. Because SMS is so widely used, everyone may now easily use the newest technology for convenient communication. From the perspective of SMS, the provision of resources that increase the use of the countless SMS initiatives is becoming more and more cost-effective. Despite the fact that SMS messages are not free, the volume of SMS spam is high because filtering technologies are always improving to reduce spam and provide the desired messages [1]. Because there are so many unnecessary messages that are of no use to the recipients and take up so much storage space, there is a rising problem with spam SMS in society. Anti-unsolicited mail learned that the initial nuances in SMS messages were created by manipulating the texts. SMS spam is a major contributor to end users' frustration with how much of their cell device's usable capital they are using, and in certain

 DOI : 10.35291/2454-9150.2023.0115

organizations, even the recipient gets compensated for receiving the SMS. Spam in SMS irritates the recipient, efficiently utilizes the phone, and in some networks, even costs the recipient for receiving the SMS. It is essential to eliminate these spam mails as a result [2]. With the help of fact samples, spam filters deceive recipients on a few different levels. When phone users cannot or do not want to communicate with one another, SMS is used as an alternative to voice calls. The problem with spam SMS is that it sends out unwanted and pointless messages that are not helpful for the customers' content traffic. The phone's storage and processing power are used by it. The method described here relies solely on automated testing, and its simplicity comes from the fact that it only requires literature review is carried out for the proposed objective, and the observations/findings are presented here with different approaches.

II. REVIEW WITH DIFFERENT ASPECTS

While doing or, existing work done by various authors is considered. In the lit rev, it is important to study the existing work done with respect to some aspects such as: database used, contents, size of the database as database(s) used is a fundamental aspect in any study research. Often, result of the proposed work depends on the databases taken for study. Some standard databases are available, and some are borrowed as per the requirement. It is observed that most of the studies have referred the UCI machine learning repository database for their work undertaken, which is having 5574 sizes.

The widely used approach is of Machine Learning for achieving accuracy by applying different algorithms of Machine Learning.

Luo GuangJun, Shah Nazir, Habib Ullah Khan, and Amin Ul Haq4 (2020). For accurate identification, we suggested using machine learning-based spam detection techniques. In this method, Communications on mobile devices: ham and spam messages are classified utilising computer learning classifiers such decision trees, K-nearest neighbours (K-NN), and logistic regression (DT). The dataset used to test the approach was collected through SMS spam. To train and test the research, the dataset is divided into two groups. The experiments' findings indicated that, with a high accuracy of 99%, the classification performance of LR is superior to that of K-NN and DT [3].

S. Sheikhi, M.T. Kheirabadi, A. Bazzazi (2020) Mobile spam, SMS spam, and machine learning-based anti-spam SMS spam is a real problem for mobile users, which troubles telecom companies because it irritates their customers and costs them money. As a result, we presented a unique machine learning technique in this study for the detection of SMS spam messages. The technique was tested on a dataset of actual SMS containing more than 5,000 messages and its accuracy and F-measure metrics were assessed. In the current study A dataset of 5,574 text messages classified as spam and ham was used in this study (legitimate) It can be found in the UCI machine learning repository for free, and the findings were also tested against three recently published publications. Comparisons of the suggested approach, Rough Tree 96.71% J48 Naive Bayes 96 34% [4].

Nilam Nur Amir Sjarif (2019) TF-IDF, Random Forest, and Spam The complexity of the messages that spammers impose has made it harder to classify spam. As a result, many for the purpose of preventing spam, methods have been created. Data on SMS spam messages will be collected using Random Forest and the terms frequency-inverse document frequency (TF-IDF) algorithms in this project. The experiment shows that the Random Forest algorithm performs best, with a 97.50% accuracy rate [5].

Dr. K.Sree Ram Murthy(2020). RNN, spam detection, and machine learning. To counter, a number of different techniques are utilised, including the Bayes classifier, nearest neighbour, SVM, and neural networks. These benefits. SVM and Bayesian are two further generic methods for categorising undesired text messages. results that are most efficient. NB Multi 77.09% SVM 85.35% K-NN 85.60% RF 90.12% Ada Boost has an RNN of 91.49% [6].

K. Yadav, P. Kumaraguru SVM and Bayesian learning The method put forth in the paper exceeds other methods assessed in PA, with a classification accuracy of 97% for non-spam and around 93% for SMS spam detection.[7].

Bollam Pragna, M.RamaBai (2019) utilised technique To compare the performance metrics of the approaches we used in this research, we used Support Vector Machines (SVM). On a dataset of SMS spam collected, the algorithm we proposed had an average accuracy of 98.49% using an SVM model. The UCI Machine Learning Repository provided the data set that was used. It contains more than 5000 SMS texts that have been classified and gathered for message spam study [8].

N. Krishnaveni and V. Radha (2021). The dataset for SMS spam collection was collected from the Kaggle repository. It contains text messages and ham/spam, each with 5574 instances. Used to methods Support Vector Machine accuracy 93.02% and Naïve Bayes accuracy 94.32% [9].

Abdallah Ghourabi Mahmood A. Mahmood (2020) The experimental findings we discuss in this study demonstrate that our CNN-LSMT model outperforms the competition. It evaluated our strategy with a very good accuracy of 98.37% [10].

Pavas Navaney Gaurav Dubey (2018) Our SVM model performs better than the naïve bayes model, properly classifying spam and ham with accuracy rates of 96.4% and 98.4%, respectively, for a total accuracy of 97.4%, as shown in the cross table [11].

 DOI : 10.35291/2454-9150.2023.0115

Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal and Pulikt Mendiratta There are 5574 English text messages in this collection. both spam and not. We conducted contrasts between 8 distinct classifiers. According to the findings of our examination of the classifiers, convolutional neural networks classifiers had the best accuracy for the two datasets, at 99.19% and 98.25 percent, and an AR value of 0.9926 and 0.9994 [12].

S. Nyamathulla1, Polavarapu Umesh2, Batchu Rudra Naga Satya Venkat3challa Divya kumar4. (2022) Through the UCI repository, we acquired a dataset. There are 5572 rows and columns in the dataset. Later, a large number of machine learning classifiers, such as naive Bayes random forest decision trees and others, were created. The deep learning model LSTM was also utilised in our research projects. To categorise the data, a variety of classification techniques were applied. The LSTM model achieved a satisfactory accuracy rating of 98.5 percent. The experiments' findings demonstrate that our methodology is more accurate than more traditional ways in detecting spam [13].

III CONCLUSION

Overall, it is observed that most of other existing systems have used UCI machine learning repository database like R and python tools and found with good accuracy. We selected a total of 13 research publications for this paper, assessed the approaches they proposed, and looked at the evaluation processes. We illustrated the information from the publicly accessible dataset that is a prerequisite for a spam filtering algorithm. We also talked about this topic's history. We have covered the search and selection process, in the studied published papers, journals, and conferences, where the research were presented in the systematic literature review. The are more work is lacking in finding the best accuracy and to do this, working on large database of Spam and Ham SMS using different techniques are required.

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my respected guide “Dr. Manasi Ram Baheti” for her able guidance and support in completing this paper. I would also like to extend my gratitude to the Head of Department respected “Prof. Dr. Sachin. N. Deshmukh” for providing me all the facilities that are required.

REFERENCES

[1] K. Yadav, S. K. Saha, P. Kumaraguru, and R. Kumra, “Take control of your smses: Designing an usable spam sms filtering system,” in 2012 IEEE 13th International Conference on Mobile Data Management. IEEE, 2012, pp. 352–355.

[2] S. J. Warade, P. A. Tijare, and S. N. Sawalkar, “An approach for sms spam detection.

[3] Luo GuangJun, Shah Nazir, Habib Ullah Khan, and Amin Ul Haq4. Received 14 May 2020; Revised 4 June 2020; Accepted 6 June 2020.

[4] Sheikhi, M.T. Kheirabadi, A. Bazzazi (2020) IJE TRANSACTIONS B: Applications Vol. 33, No. 2, (February 2020) 221-228.

[5] Nilam Nur Amir Sjarif*, Nurulhuda Firdaus Mohd Azmi, Suriayati Chuprat, Haslina Md Sarkan, Yazriwati Yahya, Suriani Mohd Sam, et al. / Procedia Computer Science 161 (2019) 509–515.

[6] Dr. K. Sree Ram Murthy1, Dr.K.Kranthi Kumar2, K.Srikar3, CH.Nithya4, K. Mahender Reddy5, K. Sai Kumar6, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056Volume: 07 Issue: 05 | May 2020.

[7] Kuldeep Yadav, Ponnurangam Kumaraguru, Atul Goyal, Ashish Gupta, and Vinayak Naik Indraprastha Institute of Information Technology (IIIT), Delhi. New Delhi, India.

[8] Bollam Pragna, M. RamaBai. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.

[9] N Krishnaveni and v Radha: Comparison of naïve bayes and SVM classifiers for Detection of Spam SMS using natural language processing DOI: 10.21917/ijsc.2021.0323.

[10] Abdallah Ghourabi Mahmood A. Mahmood (2020). Future Internet 2020,12,156; doi:10.3390/fi12090156.

[11] Pavas Navaney Gaurav Dubey (2018) 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence).

[12] Mehul Gupta, Aditya Bakliwal, Shubhangi Agarwal & Pulkit Mehndiratta Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August 2018, Noida, India.

[13] S. Nyamathulla1, Polavarapu Umesh2, Batchu Rudra Naga Satya Venkat3challa Divya kumar4 Journal of Positive School Psychology 2022, Vol. 6, No. 5, 7006–7013.

 DOI : 10.35291/2454-9150.2023.0115