A Review on Machine Learning Models for Mobile Text Spam Classification

Ramlal Agrawal[1] Prof. Pankaj Raghuwanshi[2]

***Abstract*— Spam messages today are far more sophisticated than they used to be. Attackers frequently modify message structures, incorporate abbreviations, and use social engineering techniques to deceive users. Rule-based filters struggle to keep up with this evolving nature of spam—every new variation demands manual updates. Machine learning models outperform rule-based systems by learning contextual and statistical features from vast datasets, enabling them to detect even unseen patterns of spam. This adaptability makes these models essential for robust spam control. This paper presents a comprehensive survey on redirection spamming attack detection using artificial intelligence based approaches so as to thwart spamming attacks for time critical applications. Various models have been discussed with their pros and cons.**

Keywords—Semantic Analysis, Spam Classification, Machine Learning, Deep Learning, Classification Accuracy.

I. INTRODUCTION

With the exponential growth of digital communication technologies, text messaging has become a primary medium for personal and business interactions. However, this rise has been accompanied by significant exploitation from malicious actors who disseminate spam messages to defraud users or push unwanted promotional content. Traditional rule-based spam filters are no longer sufficient to handle the scale and complexity of modern spam. This has led to the increasing reliance on machine learning models that possess the ability to learn patterns from data, adapt to new threats, and offer accurate and efficient spam detection [1].

Generally it is difficult to classify based on auto-redirect or auto-refresh tag because when web servers are heavily loaded, they may introduce such measures to release load and avoid web server crashes. Hence it becomes mandatory to look for techniques which can classify with high accuracy in time critical aspects and situations. The following section presents the basics of artificial neural networks used for spam classification [2].

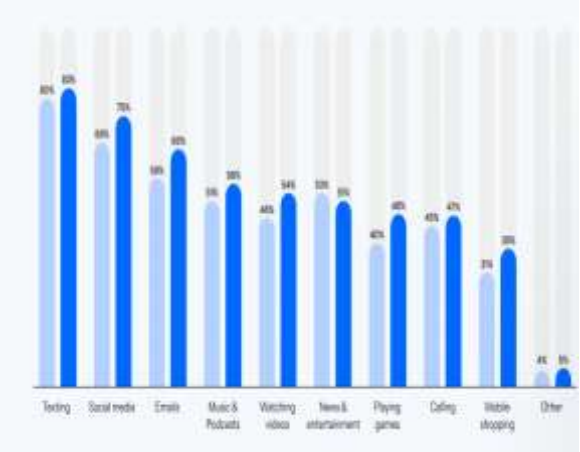

Fig.1 Mobile Activities Prone to Spamming (Source: https://simpletexting.com/blog/2025-texting-and-sms-marketing-statistics/)

Figure 1 depicts the activities most prone to spamming. Some of the spamming attacks may be benign while others may be malignant trying to redirect mobile users to malicious websites where user security may be compromised. Since the amount of data is staggering large and complex, off late machine learning based approaches are becoming common to filter out spams. One of the challenges which machine learning based approaches face for mobile spamming platforms is the limited computational and processing capabilities of hand held mobile devices [3]. This makes is necessary to design and test algorithms which are compatible with various versions of mobile operating systems and also supported by limited memory and processing hardware as there exists a lot of diversity in the mobile hardware of different devices.

II. MACHINE LEARNING MODELS FOR SPAM CLASSIFICATION

Machine learning-based spam filtering improves user safety and trust in communication systems. Spam messages often contain phishing links, malware downloads, or fraudulent financial requests. By detecting harmful content in real-time through classification models like Naive Bayes, Support Vector Machines, Random Forest, or more advanced deep learning architectures such as LSTMs and Transformers, machine learning systems offer a proactive defense mechanism against cybercrime. This ensures secure user experience and helps service providers comply with data protection and cybersecurity regulations [4].

Another major advantage of machine learning models is their ability to handle large-scale data efficiently. With billions of messages transmitted daily across the globe, manual supervision or static rules cannot ensure effective filtering. Machine learning algorithms can process vast datasets, extracting relevant linguistic features such as lexical cues, message structure, and metadata. These insights enable highly accurate predictions of whether a message is spam or legitimate (ham), reducing false positives and maintaining smooth communication flow [5].

Additionally, machine learning models continuously improve through feedback loops. As new spam patterns emerge, user reports and data collection help retrain models to maintain high precision and recall. Natural language processing techniques make detection more intelligent by considering semantics, sentiment, and intent rather than relying solely on keyword spotting. This dynamic learning capability provides a long-term, scalable, and cost-effective solution for text spam filtering [6].

Some of the most common machine learning models used are presented next:

Support Vector Machine (SVM): Before the advent of deep learning, traditional machine learning models such as Support Vector Machines (SVM), Decision Trees, Random Forests, and K-Nearest Neighbors (KNN) were widely used for classification [7]

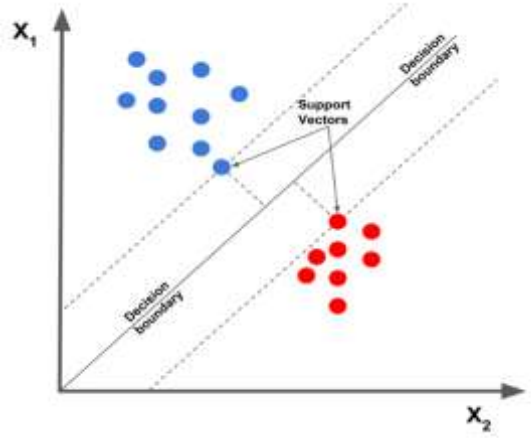

Fig.2 The SVM Model

Figure 2 depicts the SVM Model.

The SVM classifies based on the hyperplane.
The selection of the hyperplane H is done on the basis of the maximum value or separation in the Euclidean distance d given by:

$$\boldsymbol{d = \sqrt{x_1^2 + \cdots \ldots \ldots . x_n^2}} \quad \textbf{(1)}$$

Here,
x represents the separation of a sample space variables or features of the data vector,
n is the total number of such variables
d is the Euclidean distance

The (n-1) dimensional hyperplane classifies the data into categories based on the maximum separation. For a classification into one of **'m'** categories, the hyperplane lies at the maximum separation of the data vector **'X'.** The categorization of a new sample **'z'** is done based on the inequality [8]:

$$\boldsymbol{d_x^z = Min(d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z)} \quad \textbf{(2)}$$

Here,
d_x^z is the minimum separation of a new data sample from **'m'** separate categories
$d_{C1}^z, d_{C2}^z \ldots d_{C2=m}^z$ are the Euclidean distances of the new data sample **'z'** from m separate data categories.
For instance, SVMs are effective for binary classification tasks, such as distinguishing between urban and rural areas, while Random Forests are used for multi-class classification problems, such as land cover mapping. However, these models struggle with complex patterns require extensive feature engineering, which limits their scalability and accuracy

Neural Networks:

Owing to the need of non-linearity in the separation of data classes, one of the most powerful classifiers which have become popular is the artificial neural network (ANN). The neural networks are capable to implement non-linear classification along with steep learning rates. The neural network tries to emulate the human brain's functioning based on the fact that it can process parallel data streams and can learn and adapt as the data changes. This is done through the updates in the weights and activation functions [9].

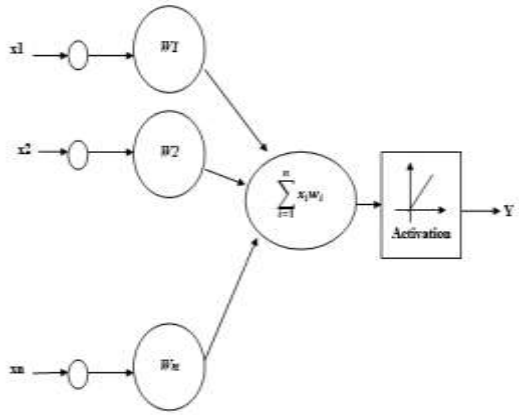

Fig.3 The ANN Model

The input-output relation of a CNN is given by:

$$y = f(\sum_{i=1}^{n} x_i w_i + b) \quad (3)$$

Here,
x denote the parallel inputs
y represents the output
w represents the bias
f represents the activation function

The neural network is a connection of such artificial neurons which are connected or stacked with each other as layers. The neural networks can be used for both regression and classification problems based on the type of data that is fed to them. Typically the neural networks have 3 major conceptual layers which are the input layer, hidden layer and output layer. The parallel inputs are fed to the input layer whose output is fed to the hidden layer. The hidden layer is responsible for analysing the data, and the output of the hidden layer goes to the output layer. The number of hidden layers depends on the nature of the dataset and problem under consideration. If the neural network has multiple hidden layers, then such a neural network is termed as a deep neural network. The training algorithm for such a deep neural network is often termed as deep learning which is a subset of machine learning. Typically, the multiple hidden layers are responsible for computation of different levels of features of the data.

Convolutional Neural Networks (CNNs): The family of CNNs are the backbone of modern satellite object detection. CNNs automatically learn hierarchical features from raw images, eliminating the need for manual feature extraction. The Convolutional Neural Networks (CNNs) can automatically extract hierarchical characteristics from images, they have become the mainstay for image classification applications. These neural networks perform exceptionally well in applications like picture identification because they are specifically made for processing organised grid data.

Convolutional, pooling, and fully linked layers are among the layers that make up a CNN's architecture. Convolutional layers identify patterns in the input image by applying filters, hence identifying local features. By reducing spatial dimensions, pooling layers preserve significant information. High-level features are integrated for categorization in fully connected layers.

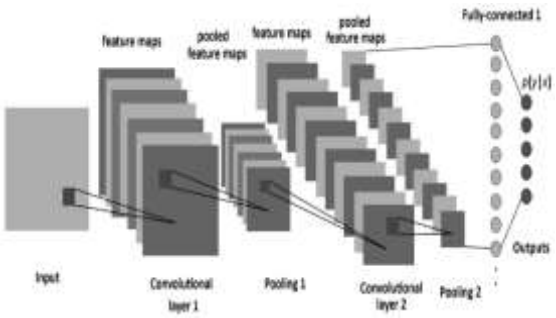

Fig.5 The CNN Model

The convolution operation is given by [10]:

$$x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \quad (4)$$

Here,
x(t) is the input
h(t) is the system under consideration.
y is the output
*is the convolution operation in continuous domain
For a discrete or digital counterpart of the data sequence, the convolution is computed using:

$$y(n) = \sum_{-\infty}^{\infty} x(k)h(n-k) \quad (5)$$

Here
x(n) is the input
h(n) is the system under consideration.
y is the output
*is the convolution operation in discrete domain

III. RELATED WORK

Various approaches have been devised for mobile spam classification.

Salman et al. [11] present present a new SMS dataset comprising more than 68K SMS messages with 61%

legitimate (ham) SMS and 39% spam messages. Notably, this dataset, authors release for further research, represents the largest publicly available SMS spam dataset to date. Authors extract semantic and syntactic features to evaluate and compare the performance of well-known machine learning based SMS spam detection methods, ranging from shallow machine learning approaches to advanced deep neural networks. They also investigate the robustness of existing SMS spam detection models and popular anti-spam services against spammers' evasion techniques. Experimental findings reveal that the majority of shallow machine learning based techniques and anti-spam services exhibit inadequate performance when it comes to accurately classifying SMS spam messages.

Agarwal et al. [12] proposed a novel approach for spam detection using natural language processing. The proposed strategy utilizes a least-squares model to modify themes and incorporates gradient descent and altering least-squares (i.e., AMALS) models for estimating missing data. TF-IDF and uniform-distribution methods perform the estimation. The performance evaluation reveals that the suggested technique exhibits a superior performance of 98% compared to the existing industry TF-IDF model in accurately predicting spam within big data ecosystems. By this model, the environment of an organization or a company can be saved from spamming or other attacks, which can lead to extracting their data for unauthorized users to protect the details.

Kabbi et al. [13] proposed that spam is a pervasive issue that affects millions worldwide, leading to significant inconvenience, time wastage, and potential financial scams. This paper proposes a novel approach to SMS spam detection involving five steps: preprocessing, feature extraction, feature fusion, feature selection, and classification. Our model is designed to simultaneously capture local, temporal, and global text message features using a hybrid deep learning model to enhance feature representation. We evaluated our model using the UCI dataset, comparing it with traditional and deep learning algorithms such as RF and BERT using cross-validation to ensure the robustness of our results. Our proposed method exhibited superior performance, achieving a good accuracy of 99.56%, surpassing other methods. The effectiveness of this method in SMS spam detection proved its potential for real-world implementation, where it could substantially mitigate the prevalence and impact of SMS spam.

Joseph et al. [14] proposed a a comparative analysis of these popular word embedding techniques for SMS spam detection by evaluating their performance on a publicly available ham and spam dataset. Authors investigate the performance of the word embedding techniques using 5 different machine learning classifiers i.e. Multinomial Naive Bayes (MNB), KNN, SVM, Random Forest and Extra Trees. Based on the dataset employed in the study, N-gram, BOW and TF-IDF with oversampling recorded the highest F1 scores of 0.99 for ham and 0.94 for spam.

Jain et al. [15] proposed an approach for the detection of spam messages. We have identified an effective feature set for text messages which classify the messages into spam or ham with high accuracy. The feature selection procedure is implemented on normalized text messages to obtain a feature vector for each message. The feature vector obtained is tested on a set of machine learning algorithms to observe their efficiency.

Adewole et al. [16] proposed a unified framework is proposed for both spam message and spam account detection tasks. Authors utilized four datasets in this study, two of which are from SMS spam message domain and the remaining two from Twitter microblog. To identify a minimal number of features for spam account detection on Twitter, this paper studied bio-inspired evolutionary search method. Using evolutionary search algorithm, a compact model for spam account detection is proposed, which is incorporated in the machine learning phase of the unified framework. The results of the various experiments conducted indicate that the proposed framework is promising for detecting both spam message and spam account with a minimal number of features.

Barushka et al. [17] proposed a technique based on integrated distribution-based balancing approach for spam classification. The concept of deep neural networks is used in this paper. The major advantage of this approach is the distribution mechanism makes the computation of different parameters for classification simpler. Deep learning makes the classification accuracy higher.

Sedhai et al. [18] proposed a technique that used semi-supervised approach for spam redirection classification mechanism. The concept used the training rules to be governed by supervised learning with an adaptive weight changing mechanism. However, the approach had the liberty of letting the weight adaptation fall into the purview of the training algorithm used.

Chen et al. [19] proposed a technique for the classification of drifted twitter spam based on statistical feature based classification. The major issues addressed in this paper, were the use of statistical features for spam classification. Drifted spam is often the result of several attached web links leading to the drifting mechanism of the tweets in social media applications with malicious URLs that can cause the spamming attacks on the web mails.

Mirza et al. [20] proposed a technique for spam classification based on hybrid feature selection. The major advantage of this approach was the fact that the hybrid parameters can be an amalgamation of both textual features and non-textual features. The evaluation of the performance of the proposed system was done on the basis of mean square error, hit rate and the accuracy. The performance of hybrid feature selection was shown to be better than the average features computation algorithms.

IV. CLASSIFICATION AND PERFORMACNE METRICS

The need for probabilistic classifiers arise from the fact that the classification problem often encounters data sets with overlapping vectors. The major challenges in spam classification are:

1) It is very difficult to detect malicious redirections because redirections are also made intentionally for non-harmful purposes like load balancing [21].

2) If successful redirection is not employed, then Web Server may crash in case requests received becomes much more than request handling capacity.

3) It is very difficult to actually detect a malicious spam and differentiate it from a load balancing redirection.

The feature selection mechanism is also important for the computation of the various parameters that include the mean square error and accuracy. However, the addition of features makes the accuracy increase at times but also increases the complexity of the training.

4) Moreover, general machine learning techniques for spam classification are prone to poisoning attacks.

Depending on the implementation, Bayesian spam filtering may be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering. A spammer practicing Bayesian poisoning will send out emails with large amounts of legitimate text (gathered from legitimate news or literary sources).

Spammer tactics include insertion of random innocuous words that are not normally associated with spam, thereby decreasing the email's spam score, making it more likely to slip past a Bayesian spam filter. However, with (for example) Paul Graham's scheme only the most significant probabilities are used, so that padding the text out with non-spam-related words does not affect the detection probability significantly. Words that normally appear in large quantities in spam may also be transformed by spammers. For example, «Viagra» would be replaced with «Viaagra» or «V!agra» in the spam message. The recipient of the message can still read the changed words, but each of these words is met more rarely by the Bayesian filter, which hinders its learning process. As a general rule, this spamming technique does not work very well, because the derived words end up recognized by the filter just like the normal ones.

The overlapping vectors make its challenging to find a clear boundary for the classification problem and often there exists only a fuzzy or non-clear boundary to demarcate among the data classes. In such overlapping classes, the final categorization of a new data vector 'X' is done based on the maximum mutual probability given by [22]:

$$P(X) = Max\{\frac{X1}{U}.\frac{X2}{U}.\ldots..\frac{Xn}{U}) \quad (6)$$

Here,
X1, X2….Xn are the multiple classes
U is the universal set containing all the classes
P(X) is the maximum probability of a data sample to belong to a particular category.

The final classification accuracy is computed as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Here.
TP represents true positive
TN represents true negative
FP represents false positive
FN represents false negative

CONCLUSION: The need for machine learning models in spam detection is driven by the rapidly evolving nature of spam attacks, the demand for automated and scalable solutions, and the necessity to protect users from cyber threats. With advancements in AI and NLP, machine learning-based systems offer a strong defense against malicious communications and contribute to a safer digital ecosystem. As messaging technologies further expand with IoT and

5G, the integration of advanced machine learning models will continue to be indispensable for identifying and mitigating text spam effectively. spam classification is a non-trivial task based on the amount and the complexity of data mobile and web servers receive in real time situations. It can be inferred from the discussions made so far that AI and ML based approaches are appropriate to cater to the needs of the web services. However, the challenging aspect in spam classification remains the accuracy that needs to be met for real life applications which may be challenging.

References

1. Liu, Xiaoxu, Haoye Lu, and Amiya Nayak. "A Spam Transformer Model for SMS Spam Detection." IEEE Access, vol. 9, 2021, pp. 80253–80263
2. Oswald, Christopher, S. E. Simon, and Anupam Bhattacharya. "SpotSpam: Intention Analysis–Driven SMS Spam Detection Using BERT Embeddings." ACM Transactions on the Web, vol. 16, no. 3, 2022, pp. 1–27
3. Altunay, Hakan C., and Zafer Albayrak. "SMS Spam Detection System Based on Deep Learning Architectures for Turkish and English Messages." Applied Sciences, vol. 14, no. 24, 2024, article 11804
4. Ghourabi, A., A. M. Mahmood, and M. Q. Alzubi. "Enhancing Spam Message Classification and Detection with Pre-trained Transformers and Ensemble Learning." Future Internet (special issues 2023–2024)
5. Shaaban, Mai A., Yasser F. Hassan, and Shawkat K. Guirguis. "Deep Convolutional Forest: A Dynamic Deep Ensemble Approach for Spam Detection in Text." arXiv, 2021. https://arxiv.org/abs/2110.15718
6. N. Sharma, "A Methodological Study of SMS Spam Classification Using Machine Learning Algorithms," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-5
7. P. Manasa et al., "Tweet Spam Detection Using Machine Learning and Swarm Optimization Techniques," in IEEE Transactions on Computational Social Systems, vol. 11, no. 4, pp. 4870-4877, Aug. 2024
8. X. Liu, H. Lu and A. Nayak, "A Spam Transformer Model for SMS Spam Detection," in IEEE Access, vol. 9, pp. 80253-80263, 2021
9. M. Ketcham, T. Ganokratanaa, P. Pramkeaw and N. Chumuang, "Spam Text Detection using Machine Learning Model," 2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), Jeju, Korea, Republic of, 2023, pp. 1-6
10. Pudasaini, S. "SMS Spam Detection Using Relevance Vector Machine." Procedia Computer Science (or similar proceedings), 2023
I. Del Rosario, B. D. P. Fernandez and D. A. Padilla, "Email Spam Classification using DistilBERT," 2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Coron, Palawan, Philippines, 2023, pp. 1-6
11. M. Salman, M. Ikram and M. A. Kaafar, "Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models," in IEEE Access, 2024, vol. 12, pp. 24306-24324.
12. R. Agarwal et al., "A Novel Approach for Spam Detection Using Natural Language Processing With AMALS Models," in IEEE Access, 2024 vol. 12, pp. 124298-124313.
13. H. A. Al-Kabbi, M. -R. Feizi-Derakhshi and S. Pashazadeh, "Multi-Type Feature Extraction and Early Fusion Framework for SMS Spam Detection," in IEEE Access, 2023 vol. 11, pp. 123756-123765
14. P. Joseph and S. Y. Yerima, "A comparative study of word embedding techniques for SMS spam detection," 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), 2022, pp. 149-155
15. AK Jain, D Goel, S Agarwal, Y Singh, G.Bajaj, "Predicting Spam Messages Using Back Propagation Neural Network", Journal of Wireless Personal Communications, Springer 2021, vol. 110, pp. 403-422.
16. KS Adewole, NB Anuar, A Kamsin, "SMSAD: a framework for spam message and spam account detection", Journal of Multimedia Tools and Applications, Springer 2021, vol. 78, pp. 78, 3925–3960.
17. Aliaksandr Barushka, Petr Hajek, "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks", Springer 2018
18. Surendra Sedhai, Aixin Sun, "Semi-Supervised Spam Detection in Twitter Stream", IEEE 2018
19. Chao Chen, Yu Wang, Jun Zhang, Yang Xiang, Wanlei Zhou, Geyong Min, "Statistical Features-Based Real-Time Detection of Drifted Twitter Spam", IEEE 2017S
20. N. Mirza, B. Patil, T. Mirza and R. Auti, "Evaluating efficiency of classifier for email spam detector using hybrid feature selection approaches," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2017, pp. 735-740
21. B. C. Sulochana, B. S. Pragada, K. Lokesh and M. Venugopalan, "PySpark-Powered ML Models for Accurate Spam Detection in Messages," 2023 2nd International Conference on Futuristic Technologies (INCOFT), Belagavi, Karnataka, India, 2023, pp. 1-6.
22. A. M. Al-Zoubi, A. M. Mora and H. Faris, "A Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines," in IEEE Access, vol. 11, pp. 72250-72271, 2023