Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

"A HYBRID NAIVE BAYES–SBERT ENSEMBLE FOR ROBUST SMS SPAM AND PHISHING DETECTION"

Muhammad Ghazanfar Ullah Khan[*1], Khaliq Ahmed[2], Muhammad Khalid[3], Saima Ishaq[4], Shilpa Kumari[5], Abdul Khaliq[6]

[*1]*Department of Computer Systems Engineering, UIT University Karachi*
[2]*Department of Computer Science Nazeer Hussain University Karachi*
[3]*Department of Computer Science DHA Suffa University Karachi*
[4]*Department of Computer Science, Institute of Business Management, Karachi*
[5]*Department of Computer Science, Iqra University, Karachi*
[6]*Department of Computer Science, Institute of Business Management, Karachi*

[*1]ghanzafar.ullah@gmail.com, [2]drkhaliq.ahmed@nhu.edu.pk, [3]muhammad.khalid@dsu.edu.pk, [4]saima.ishaq@iobm.edu.pk, [5]shilpa@iqra.edu.pk, [6]khaliq@iobm.edu.pk

DOI: https://doi.org/10.5281/zenodo.18031275

Keywords
ensemble, Naïve Bayes, SMS Spam,

Article History
Received: 11 October 2025
Accepted: 21 November 2025
Published: 18 December 2025

Corresponding Author: *
Muhammad Ghazanfar Ullah Khan

Abstract

It is the case that people communicating through text messaging have caused a big increase in the number of mobile users who can be targeted by spammers and phishers. A lot of unwanted text messages also bring about difficulty in communication and at the same time, they constitute a major risk to the user's privacy and the trust in digital communication systems. Consequently, it is a very big challenge for researchers to come up with very precise and trustworthy detection mechanisms, in conjunction with the development of such mechanisms being a very difficult task. The present paper reveals a machine learning framework that integrates the use of many technologies including both classical and modern for the efficient detection of spam and phishing attacks. The method brings together the Naive Bayes (NB) classifier that is based on lexical features extracted from the message content with Sentence-BERT (SBERT) embeddings that access deeper semantic and contextual information within the messages. While taking advantage of both models, a strategy based on probabilistic averaging is used to strengthen the overall classification's robustness. The combined dataset of 9,614 English SMS and e-mails was applied to the testing of the experiment where 20% of the data was reserved for testing. The proposed ensemble model produced an impressive 96.72% accuracy as well as 95.13% precision, 87.83% recall, and 91.33% F1-score for spam and phishing detection. The increase in performance is still a lot higher than what individual classifiers can achieve, hence hybrid ensemble learning has been shown as a powerful technique in not only detecting surface-level lexical patterns but also in understanding deeper semantic context. The findings show that hybrid models have been a strong solution in the battle against detection systems for the mobile use case.

INTRODUCTION

The use of electronic mail and short message service (SMS) has become the main means of communication in personal, business, and institutional domains. Digital communication systems have experienced a huge proliferation at the same time, and thus, the number of unsolicited and harmful messages, generically called spam, has also increased to a similar extent. Spam has turned out to be not only a major nuisance but also a channel for serious security and financial exposures when used, say, in the case of fraud, identity theft, and social engineering attacks, to mention just a few. One such method, which is slowly gaining ground, is called "smishing." In this scenario, the scamsters utilize the SMS service as the main tool for their deception; they pose as trusted companies or use the victims' phone numbers and area codes to lure them into disclosing their private or bank information [1].

The recent figures reflect the massive scale of the problem. The research states that the rise in spam messages was unbelievable, soaring from about 1.27 million in September 2021 to nearly 10.89 billion in August 2022. One prediction claims the very same volume growth was at the root of the above-mentioned scams that made 2021 the most expensive year of all time in terms of costs related to spam [1]. The figures not just emphasize the constant and somehow more advanced migration of phishing and spamming campaigns but also show how urgent it is to use new and more trustworthy detection systems in order to safeguard digital communication resources. Spam detection systems used to be built on rule-based methods such as blacklisting bad senders and filtering based on keywords or patterns in the messages historically. This classical approach to security does not provide a high level of security rather it is static. Thus it cannot be easily changed along with the attackers' evolving strategies. In the present-day scenario, the spams messages are using such tricks like obfuscation, contextual manipulation, and varying the languages to bypass the filters and at the same time making it difficult for the system to identify them. The latest statistics have revealed the surprisingly large scale of the problem. The research indicates that the surge of spam messages was extraordinary, increasing from approximately 1.27 million in September 2021 up to roughly 10.89 billion in August 2022. One prediction states that the total volume of spam and scams that made 2021 the most expensive year in terms of detection costs was all through increasing the above-mentioned volume growth [1]. The figures not only reveal the persistent and ever more sophisticated movement of phishing and spam campaigns but also point out the urgent necessity of implementing new and trustworthy detection systems to secure the digital communication resources. In the past, spam detection systems were mainly dependent on the rule-based approaches which used techniques such as blacklisting the identified bad senders and heuristics built on a dictionary of words or patterns for their operation. This traditional method of protecting was not very effective in terms of security, instead, it was fixed. Therefore, it was not easy to change it in response to the attackers' changing tactics. Current spam messages are using tricks like hiding, manipulation of context, and different linguistics to get through the filters and as a result, the system has a hard time detecting them.

Literature Review

Throughout the years, researchers have experimented with a wide array of machine learning techniques for the detection and prevention of phishing and smishing attacks. The studies, both past and current, have concluded that automatic classifiers outperform manual sorting by a significant margin especially when the sorting and monitoring are done through very deceitful and ever-changing attack schemes. The discrepancies among the studies point to the different, complex and slow advances in spam and phishing detection with varied feature extraction techniques, learning algorithms and system designs being the hallmarks of this.

In Ref. [4], the authors introduce a “SmiDCA” framework that relies on a very detailed feature extraction of 39 attributes that distinguish smishing messages from the real ones. The model gives an impressive performance in classification with an accuracy percentage of 96.40% for English and 90.33% for other languages. Dimensionality reduction is applied but still, the accuracy of the system is around 96.16%, indicating its superior performance and skill in coping with high-

dimensional feature spaces. Likewise, the study in Ref. [5] shows a security architecture that aims at eliminating smartphone vulnerabilities by integrating detection modules. The architecture encompasses content analysis of the SMS messages via Naive Bayes classification, URL filtering process, application source code analysis, and APK download detection system, all combined together. The model's detection accuracy is quoted to be 96.29%, which may be perceived as an understatement, although it certainly mirrors the intense competition among the multi-layered detection techniques possessing powerful strategies like a competitive edge.

Another point of view is that the highlight on modular architectures for spam and phishing detection is already done. Joo et al. [6] push for a thorough structure that includes a message monitoring unit, content analyzer, classification engine, and a knowledge base that smoothly incorporates the attributes of spam and phishing detection at every step.

Research Gap:

Most previous works focus on either classical models or transformers individually. This study addresses the gap by creating a **robust, hybrid ensemble** that combines both approaches to enhance detection accuracy and reliability.

Dataset Description

The research was grounded on the dataset that was formed by combining two datasets that are the most popular and the easiest to access publicly, thus ensuring its reliability and timeliness in the area of spam and phishing detection. The first dataset is the UCI SMS Spam Collection, which is a standard benchmark dataset and has been widely used in studies of SMS spam detection. The dataset is composed of SMS texts that have been marked manually as either ham (legitimate) or spam. It is a suitable dataset for implementing classical spam detection with the assistance of lexical features and analyzing word frequency distributions.

Conversely, the second dataset used is the Spam v.2 dataset, which represents the second of message categories in that it not only includes the bogus and deceiving messages but also the ones that are real, like those coming from SMS and email. The Spam v.2 dataset is a modern dataset, thus it is not biased at all; it reflects the current evil practices such as using the way of the trust source to cover frauds, fraudulent websites, and even simply using the word "free" just to lure the unsuspecting reader. The dataset contains various languages, so a rigorous preprocessing step was required to keep only the messages in English and thus maintain the linguistic uniformity of the combined dataset.

After the datasets had been merged and the duplicate entries had been removed, standard text preprocessing methods were applied. The operations carried out included converting all letters to lowercase and eliminating extra spaces.

Methodology

The accuracy and reliability of SMS spam and phishing detection are enhanced by the combination of machine learning techniques which include traditional and modern methods using semantic embedding as well to be used in this study, the overall process is shown in figure 1. A system with three principal parts is suggested: The first one is a Naive Bayes classifier, the second one is a Sentence-BERT (SBERT)-based classifier, and the third one is a voting model that combines the advantages of both through probabilistic averaging...

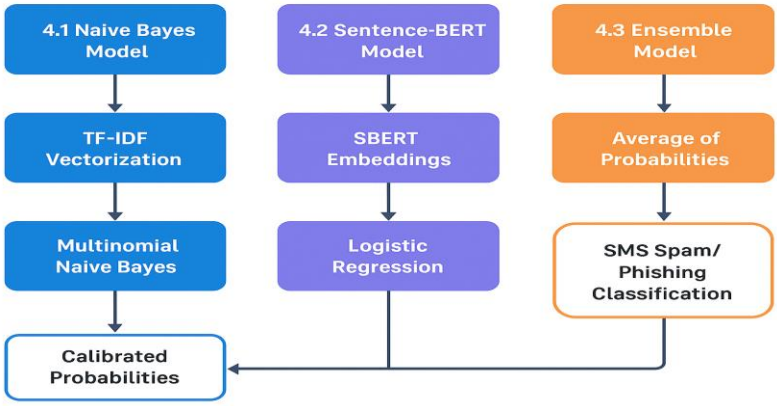

Figure 1; Flow diagram of the overall process

Naive Bayes Model

The Naive Bayes (NB) classifier was the very first one to be employed in the study as a reference model. The three reasons for this selection were: the simplicity of the model, the very good classification performance, and the reliability of the model in the text classification tasks. TF-IDF transformation of raw text input into a numerical format was the first step in the feature extraction process. After that, unigrams and bigrams were considered as part of the feature extraction process. The model learned not only the occurrence of individual words but also the pairs of words that are sometimes used as a disguise for spam or phishing messages and are thus related to them to some extent.

In general, the research was about making the models less specific by using data cleaning and removing rare words, which practically affected less than two documents, and controlling the presence of very common words by imposing a maximum document frequency limit. The TF-IDF feature matrix was used for training the Multinomial Naive Bayes classifier which is recognized for its excellent performance on discrete text data. The smoothing parameter α was chosen as 0.5, which is the point of bias-variance balance and at the same time, it avoids giving unseen words zero probability.

On the one hand, the Naive Bayes classifier can recognize the lexical patterns while on the other hand it is not always the case that its probability outputs are well-calibrated. In order to address this issue, the authors utilized cross-validation-based techniques for probability calibration. This step plays an important part in ensemble learning since it ensures that the predicted probabilities are very close to the actual class probabilities. Then, the Naive Bayes model of

Sentence-BERT Model

Naive Bayes was limited to basic lexical features only, and hence, it could not comprehend the semantic interrelations in the text. To overcome this limitation, the next part of the system introduces BERT for sentences (SBERT), which is a transformer-based model that not only produces semantically rich embeddings but also captures the semantic similarity of the input text. The pre-trained SBERT model All-MiniLM-L6-v2 was selected for its very good compromise between the computational power requirement and the quality of embedding. By means of SBERT embeddings, each and every SMS message was transformed into a dense vector of fixed length which conveys the meaning in relation to the context, shows the syntactic structure, and also demonstrates the level of similarity among the messages from the semantic angle. These embeddings are particularly suitable for short-text classification tasks such as SMS spam detection where word order and contextual hints are very important factors. A Logistic Regression model was built on the SBERT embeddings for binary classification purposes. Logistic Regression was the preferred method because of its excellent interpretability, good performance, and providing well-calibrated probability estimates as a by-product.

Class weighting was employed during the training of the model to address the issue of imbalanced classes in the dataset. The classifier generates probabilistic outputs that specify the probability of a message being in the spam or phishing group. These probabilities are treated as another input for the ensemble model.

Ensemble Model

The last step of the methodology is to create an ensemble model that combines the probabilistic outputs of two classifiers, namely Naive Bayes and SBERT-based classifier. The use of ensemble learning was meant to allow for both models to support each other; in this manner, the lexical sensitivity of Naive Bayes and the semantic understanding of SBERT could be combined and thus lead to a model with an overall better performance.

The ensemble prediction was determined as the mean of the calibrated probabilities provided by the classifiers Naive Bayes and SBERT. A decision threshold of 0.5 was set to classify the messages as either legitimate (ham) or spam/phishing. This averaging method, although simple, demonstrates great power in reducing the negative effects of the errors made by the individual models and hence increasing the generalization performance of the entire model.

The performance of the model has been evaluated through the use of various metrics such as accuracy, precision, recall, and F1 score which together give a very good indication of the quality of the classification. Furthermore, to obtain a better understanding of false positives and false negatives, a confusion matrix was made. This approach to model evaluation guarantees the full spectrum understanding of the model's capability especially when it comes to separating spam and phishing messages from the legit ones with the least misclassification of the latter group.

Experiments and Results

This section describes the experimental setup and a detailed evaluation of the performance of individual classifiers and the proposed ensemble model. The joint SMS and email dataset of 9,614 English messages was used for the experiments, which were divided into training and testing parts in an 80%–20% manner. The evaluation is carried out using primary classification metrics, which include accuracy, precision, recall, F1-score and confusion matrix analysis

Individual Model Performance

Before the construction of the ensemble was started, the single models of a TF-IDF-based Multinomial Naive Bayes classifier and a combination of a Sentence-BERT (SBERT) embedding model and Logistic Regression (LR) were analyzed. These two models represent two different ways of applying machine learning: feature-based lexical learning and learning based on semantic representation.

In this scenario, the Bayes model achieved an accuracy of 94.54%, along with a precision of 93.06% and a recall of 78.04%, an F1-score of 84.89%. High precision indicates that when the model categorizes a message as spam, it is usually right. In contrast, the recall being so much lower implies that the Bayes classifier is indeed neglecting a considerable part of spam or phishing messages. It can be inferred that this weakness is due to the model depending solely on the most fundamental word frequency features that are unable to correctly portray the semantic subtleties and the content of the doubtful phishing messages, thus showing the relationship of such messages with legitimate ones.

The SBERT plus Logistic Regression model, nevertheless, obtained a remarkable accuracy of 94.33%, with precision, recall, and F1-score being 80.50%, 93.92%, and 86.69%, respectively. Such a high recall proves that the model based on SBERT can indeed detect spam and phishing messages quite successfully, even if they depend on very subtle semantic clues or misleading wording. However, the precision that is not as high points out that the model has a larger number of false positives, i.e., messages that are wrongly labeled as spam despite being legitimate. This is an indication that the model is sensitive to semantic similarity, which may even lead to instances where it confuses the promotions or urgent messages as spam.

Table 1: Individual Model Performance

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	94.54%	93.06%	78.04%	84.89%
SBERT + LR	94.33%	80.50%	93.92%	86.69%

These results as given in table 1 highlight the complementary strengths and weaknesses of the two individual models. While Naive Bayes excels in precision and conservative spam labeling, SBERT prioritizes recall and aggressive detection of malicious messages.

Ensemble Performance (NB + SBERT)

To leverage the complementary characteristics of both classifiers, an ensemble model was constructed by averaging the calibrated probability outputs of the Naive Bayes and SBERT models. The ensemble achieved an overall accuracy of **96.72%**, significantly outperforming both individual models.

Table 2: Ensemble Performance (NB + SBERT)

Class	Precision	Recall	F1-score	Support
Ham (0)	0.9708	0.9890	0.9798	1545
Spam (1)	0.9513	0.8783	0.9133	378
Overall Accuracy			0.9672	1923

In terms of class-wise performance as given in table 2, the ensemble reached a precision of 0.9708, recall of 0.9890, and an F1-score of 0.9798 for the legitimate (ham) messages. The performance for spam and phishing messages was precision 0.9513, recall 0.8783, and F1-score 0.9133. The results suggest that the ensemble has managed to harmonize perfectly the very high precision of Naive Bayes with the excellent recall of SBERT.

The confusion matrix further illustrates this improvement:

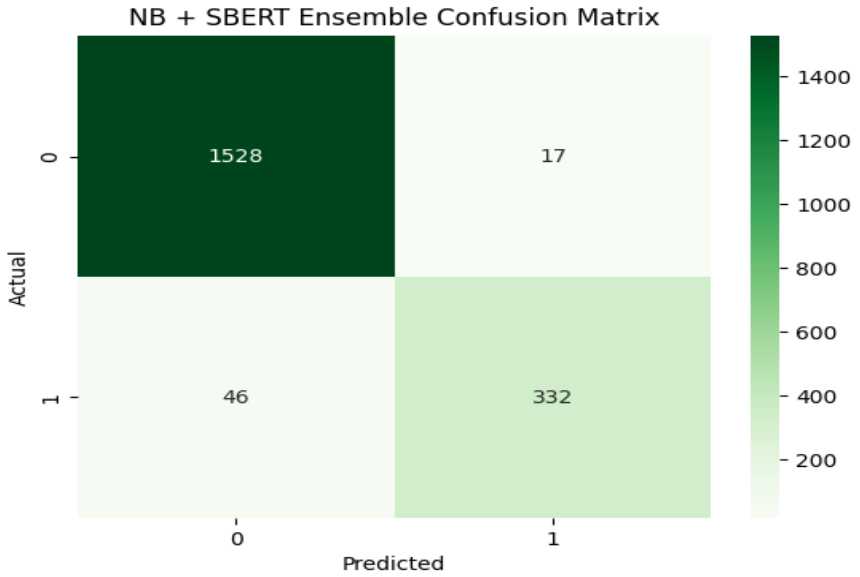

Figure 2: confusion matrix results

The false positive rate was very low since just 17 real messages were misclassified as spam as given in figure 2. On top of this, 378 spam/phishing messages were correctly identified, out of which 332 were accurately detected. This means that the system has been able to reduce the number of false negatives considerably when compared to the Naive Bayes model alone, thus demonstrating its strong detection capability

Discussion of Results

The ensemble model clearly reveals its outstanding performance by reducing both false positives and false negatives to a great extent. The very high accuracy combined with the balanced F1 score indicates that using both lexical and semantic learning schemes together is o time for SMS spam and phishing detection. The surprising aspect is that the ensemble not only provides very good ham detection accuracy, but also at the same time enhances the robustness of spam detection significantly. These findings are the confirmation of the proposed hybrid model as a reliable and practical method for deploying security systems in real-time messaging.

Discussion

- Hybrid ensembles not only capitalize on the lexical strengths of Naive Bayes but also benefit from the semantic understanding of SBERT.
- The ensemble performs with an accuracy of 96.72%, showing its ability to catch sophisticated phishing messages that were not detected by the traditional models.
- Limitations: The performance could be further enhanced with the addition of DistilBERT, but this would involve higher computational costs.

Conclusion and Future Work

The study has come up with a very good and powerful hybrid ensemble approach to detect spam and phishing in SMS by combining a traditional ML model with a modern transformer-based semantic representation method and executing them together. To be more precise, the suggested method was a combination of the TF–IDF–based Multinomial Naive Bayes classifier with Sentence-BERT (SBERT) embeddings which were further enhanced by the use of both lexical and semantic feature learning. The experiments on a merged data set of 9,614 English messages demonstrated the superiority of the ensemble method over the separate models with an overall accuracy of 96.72% and a strong precision-recall balance during the whole evaluation.

The results endorse the notion that the classic classifiers like Naive Bayes are still among the most appreciated ones for text sorting because of their simplicity, clarity, and overall low computation cost. In contrast, SBERT, with its powerful context and sense capturing ability, is far superior to the case of the frequency-based techniques that have been this aspect ignored. The implementation of the two methods through probabilistic ensemble learning resulted in the proposed system being able to greatly reduce the occurrence of false positives as well as false negatives, hence more reliable spam and phishing detection. This tradeoff is, in fact, very important in real-life situations where the consequences of either mistakenly blocking a legitimate message or letting a malicious one through can be very severe.

Notwithstanding the fact that the proposed system exhibited a superb performance, there are several areas still left for future research to explore. To start off, the trial may practice the fusion of DistilBERT or alternative miniature transformer models to enhance referring semantic while at the same time being computationally economical and effective. Besides, increasing the size of the dataset through the inclusion of multilingual messages would provide an insight into the model's performance across various language groups.

REFERENCES

1. Slicktext. 17 spam text statistics & spam text examples for 2023. Accessed 25 December 2024. Available from: https://www.slicktext.com/blog/2022/10/17-spam-text-statisitics-for-2022/
2. Tarek MM, Abd-El-Hafeez T. Developing an efficient method for automatic threshold detection based on hybrid feature selection approach. In: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics. Springer; 2020. p. 45-54.
3. Girgis MR, Mahmoud TM, Abd-El-Hafeez T. A new effective system for filtering pornography videos. Int J Comput Sci Eng. 2010; 2(9): 123-30.
4. Sonowal G, Kuppusamy KS. Smidca: an anti-smishing model with machine learning approach. The Comput J. 2018; 61(8): 1143-57. doi: 10.1093/comjnl/bxy039.
5. Mishra S, Soni D. Smishing detector: a security model to detect smishing through sms content analysis and url behavior analysis. Future Gener Comput Syst. 2020; 108: 803-15. doi: 10.1016/ j.future.2020.03.021.
6. Woong Joo J, Moon SY, Singh S, Park JH. S-detector: an enhanced security model for detecting smishing attack for mobile computing. Telecommun Syst. 2017; 66(1): 29-38. doi: 10.1007/ s11235-016-0269-9.
7. Ghourabi A, Mahmood MA, Alzubi QM. A hybrid cnn-lstm model for sms spam detection in Arabic and English messages. Future Internet. 2020; 12(9): 156. doi: 10.3390/fi12090156.
8. Roy PK, Singh JP, Banerjee S. Deep learning to filter sms spam. Future Gener Comput Syst. 2020; 102: 524-33. doi: 10.1016/j.future.2019.09.001.
9. Xia T, Chen X. A weighted feature enhanced hidden markov model for spam sms filtering. Neurocomputing. 2021; 444: 48-58. doi: 10.1016/j.neucom.2021.02.075. 10. Liu X, Lu H, Nayak A. A spam transformer model for sms spam detection. IEEE Access. 2021; 9: 80253-63. doi: 10.1109/access.2021.3081479.
11. Gupta A, Patil J, Soni S, Rajan A. Email spam detection using multi head cnn bigru network. In: International Conference on Advanced Network Technologies and Intelligent Computing. Springer; 2022. p. 29-46.
12. Silpa C, Niya Mirza S, Prathyusha S, Latha Reddy PNS, Hrudaya UJ, Vivek M. A meta classifier model for sms spam detection using multinomialnb linearsvc algorithms. In: 2023 International Conference on Networking and Communications (ICNWC). IEEE; 2023. p. 1-6.
13. Wanda P. Gruspam: robust e mail spam detection using gated recurrent unit (gru) algorithm. Int J Inf Technol. 2023; 15(8): 4315-22. doi: 10.1007/s41870-023-01516-z.

14. Mahmoud TM, Abd-El-Hafeez T. A new feature selection method based on frequent and associated itemsets for text classification. Res Sq. 2021. preprint. Available from: https://www.researchsquare.com/
15. Mahmoud TM, Abd-El-Hafeez T. The effect of rebalancing techniques on the classification performance in cyberbullying datasets. Neural Comput Appl. 2023; 35: 12345-56.
16. Mahmoud TM, Abd-El-Hafeez T, Omar A. A highly efficient content-based approach to filter pornography websites. Int J Comput Appl. 2012; 50(3): 1-7.
17. Girgis MR, Mahmoud TM, Abd-El-Hafeez T. A system for extracting images and urls from web pages. Int J Comput Appl. 2013; 75(12): 25-30.
18. Zhang Z, Mahmoud TM, Abd-El-Hafeez T. Topic extraction and interactive knowledge graphs for learning resources. Sustain. 2022; 14(1): 226.
19. Anand Sharma N, Ali ABMS, Kabir MA. A review of sentiment analysis: tasks, applications, and deep learning techniques. Int J Data Sci Anal. 2025; 19(3): 351-88. doi: 10.1007/s41060-024-00594-x.
20. Lopez-Joya S, Diaz-Garcia JA, Ruiz MD, Martin-Bautista MJ. Dissecting a social bot powered by generative ai: anatomy, new trends and challenges. Soc Netw Anal Min. 2025; 15(1): 7. doi: 10.1007/s13278-025-01410-5.
21. Batiuk T, Dosyn D. Intellectual analysis of textual data in social networks using bert and xgboost. Info Syst Netw. 2025; 17: 44-60. doi: 10.23939/sisn2025.17.044.
22. Shaba Sayeed Md, Kalyan Dutta I. Detecting malicious urls in brushing scams: a machine learning approach with human-centered cybersecurity. In: 2025 IEEE World AI IoT Congress (AIIoT). IEEE; 2025. p. 0062-9.