Spam Detection Using Machine Learning

PROF. RAJENDRA ARAKH[1], ARJIT KUMAR[2], ARYAN MISHRA[3], ANSHUL SINGH PATEL[4], ASTHA SRIVAS[5]

[1, 2, 3, 4, 5] *Department of Computer Science & Engineering, Shri Ram Institute of Technology Jabalpur*

Abstract— Email spam is a growing problem, causing frustration for users and posing risks to their security. To combat this issue, researchers have turned to machine learning techniques like Naive Bayes. This study compares Naive Bayes with other methods like Support Vector Machine (SVM) to see which is better at spotting spam. Using datasets of spam and legitimate emails, the researchers tested Naive Bayes and SVM. They found that while both methods were effective, SVM had slightly higher accuracy, with Naive Bayes close behind. The study also discusses the challenges of spam detection and the importance of machine learning in addressing this issue. By comparing different methods, it provides valuable insights into how we can better protect against email spam.

Index Terms— Spam detection, Machine learning, Naive Bayes, Support Vector Machine, Email security, Classification

I. INTRODUCTION

SMS (Short Message Service) has become a widely used communication method due to its quick responses, accessibility, and cost-effectiveness. However, it faces the challenge of spam messages, which are unsolicited commercial messages that disrupt users' experience by slowing down devices, causing storage issues, and invading privacy. Various techniques, such as blacklisting, naive bluesman, and keyword-matching algorithms, are used to identify and mitigate spam. Unfortunately, spam messages not only inconvenience users but also pose security risks, as they can be utilized by cybercriminals and advertisers alike. With the prevalence of spam, privacy becomes a concern, as responding to these messages can compromise personal information. Research has shown that millions of mobile users receive spam SMS daily, highlighting the urgency of addressing this issue.

Spam messages contrast with "ham" messages, which are legitimate and desired communications. While spam has negative effects on device performance and storage, spam messages are those that users welcome and expect to receive. However, despite efforts to combat spam, it remains a persistent problem in personal communication channels. To effectively tackle spam, businesses and researchers are developing various filtering techniques. These methods range from standard spam filtering, which employs rules and protocols to classify messages, to more sophisticated approaches like client-based filtering and enterprise-grade spam filtering. Additionally, case-based filtering, a traditional machine learning technique, categorizes emails into spam and genuine categories based on collected data and vector expressions. Overall, the challenge of SMS spam necessitates ongoing research and innovation to develop effective filtering methods that safeguard users' communication experiences and privacy.

1.1 What Is Spam?

Unwanted and unpleasant text messages in the form of spam are those that we repeatedly get via transmission channels. Spam messages have an impact on a device's performance, power, and storage system. In short, spam has proven to be the most unpleasant aspect of personal communication.

1.2 What Is Ham?

Ham refers to messages that we receive from end devices that are not spam and are on a good list of requested and wanted messages. About 2001, Spam Bayes first used the term "ham," which is currently recognized to mean "e-mail and messages that are commonly appreciated and aren't deemed spam. Its utilization is especially normal among antispam software developers, and not broadly known somewhere else; as a general rule, it is messages implemented to use allowance and task filters.

Spam filter techniques:

Spam filter techniques: Spam emails are becoming more and more prevalent in politics, education, chain messaging, stock market recommendations, and marketing. For effective spam identification and filtering, numerous businesses are currently developing various methods and algorithms. To comprehend the filtering process, we discuss a few filtering mechanisms in this part.

1. The Common Spam Filtering Technique

A filtering system that employs a set of rules and uses those set of protocols as a classifier is known as standard spam filtering. The first phase is the implementation of content filters, which identify spam using artificial intelligence methods. The second phase involves the implementation of the email header filter, which extracts the header data from the email. After that, backlist filters are applied to the emails to weed out spam emails by securing the emails originating from the backlist file. The next step is the implementation of rule-based filters, which identify the sender based on the subject line and user- defined characteristics. Finally, a technique that enables the account holder to send

2. Filtering of Spam on the Client Side

A client is a person who has access to an email network or the Internet and can send or receive emails. Several rules and procedures for ensuring secure communications transmission between persons and organizations are offered by spam detection at the client point. A client needs to install various working frameworks on his or her system for data transmission. By connecting with client mail agents and composing, receiving, and handling the incoming emails, such systems filter the client's mailbox

3. Commercial-Grade Spam Filtering

The process of detecting email spam at the enterprise level involves installing different filtering frameworks on the server, interacting with the mail transfer agent, and categorizing the gathered emails as either spam or ham. This system client employs the system regularly and successfully on a network where emails are filtered using an enterprise filtering technique. The rule of ranking the email is used by existing spam detection techniques. This principle specifies a ranking function and generates a score for each post. A certain score or rating is assigned to the spam or ham message. Since spammers employ various strategies, all jobs are routinely adjusted by adding a list-based technique to automatically block the messages

4. Spam Filtering Using Cases

The case-based or sample-based spam filtering system is one of the well-known and traditional machine learning techniques for spam detection.

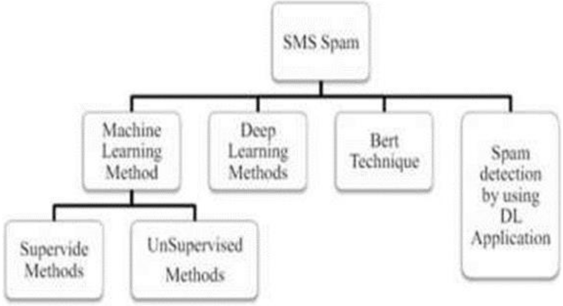


Fig. 1. Machine Learning Techniques

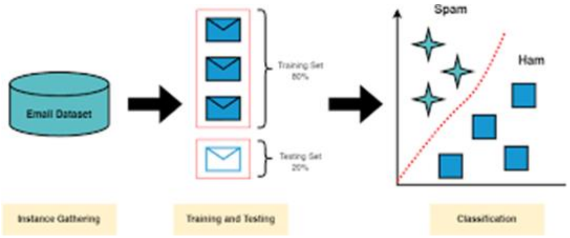


Fig. 2. Approaches to Filter Spam

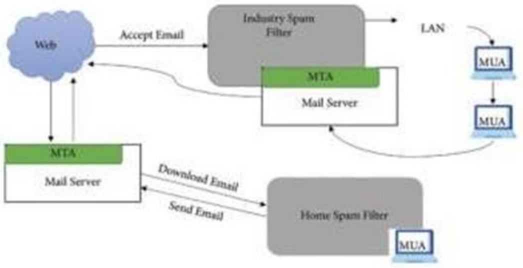


Fig 3. Client based and Enterprise based filtering

II. LITERATURE REVIEW

Spam dispatch bracket is an evolving and challenging problem, and numerous machine- literacy ways have been extensively explored to ameliorate its perfection and delicacy. Several once studies have delved different aspects of spam dispatch bracket, including the operation of machine literacy approaches, inimical approaches, the use of ensemble styles, and unsupervised literacy. Nikhil Kumar etal.'s 2020 study handed a discrepancy of colorful machine learning algorithms in the field of spam bracket.

They used support vector classifier, K- nearest neighbor, Naive Bayes, decision tree, arbitrary timber, AdaBoost classifier and Bagging classifier. In their study, support vector classifier achieved0.92 perfection, K- nearest neighbor reached0.92, Naive Bayes attained0.87, decision tree achieved0.94, arbitrary timber scored0.90, Ada Boost classifier reached0.95, and Bagging classifier attained0.94 perfection. In our study, we employed a different dataset, and our base models demonstrated perfection values nearly aligned with their reported results, frequently surpassing0.92. Akash Junnarkar etal(2021) conducted a series of trials on Enron dataset by applying four bracket algorithms. They applied SVM, RF, NB, DT and KNN with achieved rigor as97.83,97.60,95.48,90.90 and95.29, independently. SVM surfaced as name pantomime nearly followed by arbitrary timber classifier. The authors also proposed implicit exploration direction about farther refining delicacy through the relinquishment of computationally precious yet largely precise ensemble ways like XG boost. In a study conducted by W.A Awad etal, the performance of six machine literacy styles in the environment of spam bracket was epitomized using spamassasin dataset. In terms of delicacy, for the Naïve Bayes(NB) system, delicacy stood at99.46. The SVM achieved an delicacy of , and KNN algorithm showed an delicacy of96.20. In same study, neural network(NN) approach had delicacy of96.83. The artificial vulnerable system(AIS) achieved, an delicacy of96.23. Incipiently, the rough sets(RS) system had an delicacy of97.42. In their study, Zhang etal. reviewed the inimical styles used to shirk spam dispatch bracket styles and bandied the styles proposed to fight these attacks.

They also stressed the constraints of presented styles and and suggested some guidelines for implicit exploration in the field of spam dispatch bracket. In their study published in 2020, Shaukat etal. estimated the working of colorful ML styles for spam dispatch bracket comprising DT, SVM and NB classifiers.

They observed that the support vector machines showed analogous performance to decision trees. Chensu Zhao etal. bandied ensemble literacy grounded spam discovery with imbalanced data in social networks. The miscellaneous- grounded ensemble fashion is used in the imbalance class to descry spam in OSN. The base and combine modules are integrated for chancing spam in an OSN

Nikhil Govil et al. proposed the ML- ML-grounded spam discovery medium for precluding colorful phishing attacks through dictionary generation. After generating the wordbook, the features are generated by using ML algorithms. Later, the generated features are tested completely and passed to the NB algorithm. The NB algorithm calculates the probability rate of the emails and classifies them as spam or ham. Compared to other ML algorithms, the NB gives low performance and works well for dispatch-grounded spam discovery.

Mehul Gupta etal. study spam discovery in SMS by using ML algorithms. The deep literacy- grounded convolutional neural network(CNN) works better than the SVM and NB algorithms. Likewise, image-grounded spam discovery is also done through the CNN fashion. This fashion works well for some lower datasets and increases complexity rates in large datasets.

Faiza Masood etal. descry spam and fake druggies on the social network. The malware waking system and retrogression vaticination models are used for the fake content vaticination. The Twitter content is anatomized to identify fake content and druggies, spam in the URLs, and trending motifs. This work anatomized in detail the forestallment of fake accounts and the spread of fake news. In general, fake news and stoner prognostications are extremely delicate to reuse when dealing with large quantities of media data.

Yosef Hasan Fayez Jbara etal. proposed spam discovery on Twitter using a URL- grounded discovery fashion. currently, spammers are the major platform to demand social networks and spread inapplicable data to druggies. In particular, Twitter is the most prominent network to spread spam among social networks. To avoid this spread, the author used URL and ML- grounded discovery ways. Compared to other ML algorithms, the RF- grounded bracket fashion provides a advanced delicacy rate of99.2 in this process. In this work, 70 was used as training data, and 30 was used for testing purposes

Asif Karim etal. surveyed the state of intelligent spam discovery in dispatch. Both artificial intelligence and ML styles are used for intelligent spam discovery. This combined approach defended emails from phishing attacks. piecemeal from content filtering, the other styles are less covered in this analysis.

Guang- Bin Huang etal. proposed retrogression and multiclass bracket- grounded extreme literacy ways. It shows that both the literacy frame of SVM and extreme literacy machines(ELM) can be enforced. It has better scalability and briskly learning speed. But it provides veritably low performance. Poria Pirozmand etal. used the force- grounded heuristic algorithm for OSN spam discovery. The ML and deep literacy-grounded integrated fashion is used for spam filtering in OSN. The SVM, inheritable Algorithm (GA), and Gravitational Emulation Original Hunt Algorithm(GELS) are integrated to sludge spam in OSN.

III. METHODOLOGY

Data Collection : We did a thorough search to find and compile publicly available SMS datasets. We looked through various sources like GitHub repositories, Google Scholar, and the internet using keywords like ‘‘SMS’’, ‘‘SMS messages’’, ‘‘SMS dataset’’, ‘‘Text Messages’’, and ‘‘Short Message Service’’. We filtered out the results to include only those that mentioned publicly available datasets. This effort led us to gather a collection of 179,440 SMS examples in multiple languages from different public and research sources. You can find the detailed sources in . Apart from these sources, we also looked on Twitter to find tweets that reported SMS spam. These tweets were often shared as images or screenshots. We collected such tweets from January 2012 to December 2017 and from August 2022 to July 2023, ensuring that we covered more than just the Spam Hunter dataset. Additionally, we visited scam observatory websites like Scam watch and Action Fraud to download public images and screenshots of SMS scams reported by victims.

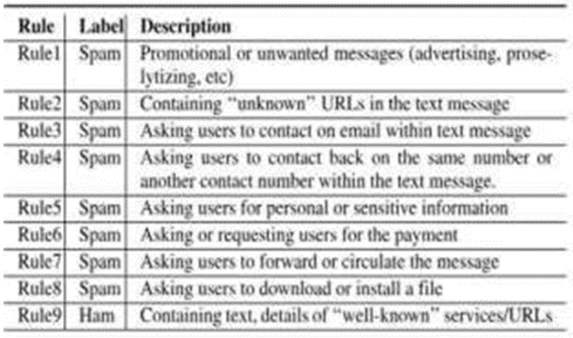
Data Augmentation:

Our main thing is to identify SMS spam dispatches written in English. To achieve this, we start by preprocessing the data, which involves removing indistinguishable andnon- English dispatches from our combined dataset. We use a two- pass filtering approach for this purpose.

In the first pass, we use Python’s langdetect library to determine the language of each SMS. Any dispatches linked asnon-English are incontinently removed. also, the remaining SMS dispatches suffer a alternate round of filtering using the Googletrans library. This step further excludes anynon-English dispatches from the dataset. It's important to note that we use the Googletrans API for this phase due to limitations on free stoner API calls. also, we convert images(screenshots of SMSes) from Twitter and fiddle lookouts into textbook using Python’s pytesseract library.

After these way, we attained a dataset of 62,114 unique English language SMSes from our consolidated data and from our collected data after removing duplicates, non- English dispatches, and labeling. Out of these, 60,032 SMSes are unlabelled , and we manually annotate them as either Spam or Ham(licit) using a set of rules outlined in . Presents nine rules that guide our homemade labeling process for SMS as Spam or Ham. These rules were developed after assaying labeled SMS corpora, conversations among the authors, and reviewing colorful fiddle types on fiddlelookouts like fiddle Watch and Action Fraud. A platoon of three experimenters collectively anatomized the unlabeled SMSes in the consolidated dataset and labeled them grounded on the defined criteria. Any disagreement in assigned markers were resolved through conversations to reach a agreement.

Tables: Rules for labelling Spam SMS in our Dataset



Rule	Label	Description
Rule1	Spam	Promotional or unwanted messages (advertising, proselytizing, etc)
Rule2	Spam	Containing "unknown" URLs in the text message
Rule3	Spam	Asking users to contact on email within text message
Rule4	Spam	Asking users to contact back on the same number or another contact number within the text message.
Rule5	Spam	Asking users for personal or sensitive information
Rule6	Spam	Asking or requesting users for the payment
Rule7	Spam	Asking users to forward or circulate the message
Rule8	Spam	Asking users to download or install a file
Rule9	Ham	Containing text, details of "well-known" services/URLs

The experimental process involves several key steps, as illustrated in Figure 4: preprocessing the combined dataset, comparing different feature models, selecting appropriate machine learning techniques, and assessing the impact of various evasion techniques on the ML models. Let's delve into these steps further.

A. Processing and Splitting Data:

We start by preprocessing the consolidated dataset, aiming to remove unnecessary characters and stop words. For this task, we employ the NLTK library. Furthermore, using the scikit-learn library, we divide the dataset into three subsets: train (80%), test (20%), and hold-out. The hold-out set, comprising 225 randomly chosen spam SMS, is specifically reserved for validation purposes messages.

This subset is used to evaluate the performance of the machine learning models, which will be elaborated in Section V-D. The train set is employed to train the ML models, while the test set is utilized to assess the model's performance on previously unseen data.

B. Feature Extraction:

Before applying machine learning models, we need to convert the SMS messages in our dataset into a structured format called a feature space. To do this, we use various techniques to transform the list of words in each message into a feature vector, which includes both syntactic (how words are arranged) and semantic (meaning of words) features. Let's explore these techniques in simpler terms.

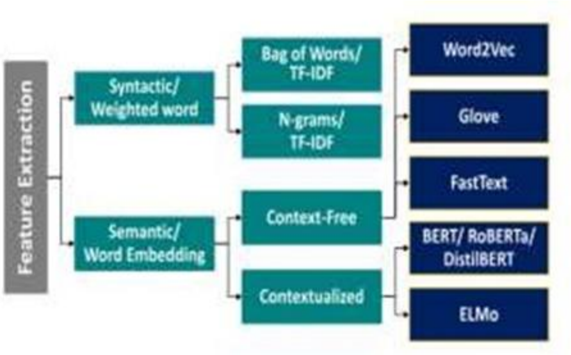


Fig. 4. Feature Extraction or Representation Techniques

1) Syntactic / Non-Semantic Count-Based Vector Space Model:

First, we convert the raw text of SMS messages into numerical features using techniques like a bag of words (BoW) and n- grams (which are sequences of adjacent words). We also use term frequency-inverse document frequency (TF-IDF) to measure the importance of words in the messages and the entire dataset. These techniques help us capture word frequency and pairs of words in each message, considering the importance of uncommon words. We implement these techniques using the scikit-learn library in Python.

2) Semantic / Word Embedding:

While syntactic representations capture the structure of words, they may not accurately represent their meaning. To address this, we use word embedding to create semantic feature vectors for each word, capturing their meaning and context.

a: Context-Independent Vector Space Model:

We use classic word embeddings like Word2Vec and GLOVE to create static representations of words, which means their meanings don't change during training. These embeddings help us understand the meaning of words in isolation.

b: Context-Dependent Vector Space Model:

To capture the contextual meaning of words, we use contextualized word embeddings like BERT and ELMo. These embeddings understand the context of words in sentences, helping us capture the nuanced meanings of words based on their surrounding words.

We generate these embeddings using Python libraries like SimpleTransformers and TensorFlow Hub.
These techniques help us extract both syntactic and semantic features from SMS messages, enabling us to better understand and analyze their content.

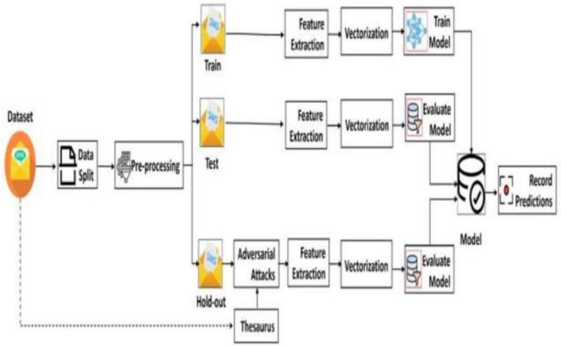


Fig. 5. Overview of Evalution Methodology

Machine Learning Algorithms:
Machine learning helps predict and classify data, a key component of Artificial Intelligence. There are two main types of machine learning methods:

1- Supervised Learning Algorithm:
In supervised learning, algorithms learn from labeled data. Various techniques are evaluated for SMS spam detection using two datasets gathered from free sources. These datasets are organized using preprocessing techniques like tokenization and TF-IDF, along with correlation algorithms and deep learning classifiers such as decision trees, Support Vector Machines (SVM), artificial neural networks (ANN), random forests, AdaBoost, Convolutional Neural Networks (CNN), and Naive Bayes (NB). SVM, NB, and entropy calculation were used to identify spam and ham messages, with SVM achieving the highest accuracy of 97.4% using a dataset of about 5574 records prepared with stop word removal and tokenization

2- Unsupervised Learning Algorithm:
Unsupervised learning involves algorithms learning from unlabeled data. Weka and RapidMiner, two different tools for arranging data, were used for spam identification. They utilize AI algorithms for clustering and classification, achieving high precision rates, with SVM in Weka reaching 99.3% accuracy in 1.54 seconds for clustering and KMeans achieving remarkable results in 2.7 seconds. RapidMiner's SVM achieved 96.64% accuracy in 21 seconds and K-Means in 37.0 seconds.

3- Deep Learning Methods:
Deep learning methods involve neural networks with multiple layers of abstraction. Convolutional neural networks (CNN) were used to classify spam and non-spam messages, achieving high accuracy rates. Additionally, a method for discretely identifying spam messages on cell phones was discussed, utilizing octet-based components and various classifiers like AdaBoost, decision trees, and K nearest neighbor (KNN) algorithms.

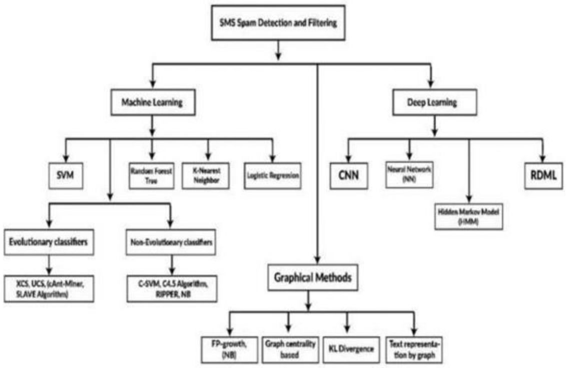


Fig. 6. Spam Detection Model

CONCLUSION

In conclusion, this study underscores the significance of effective spam discovery mechanisms in various disciplines, from dispatch communication to social media platforms. Through the operation of advanced ways analogous as the BERT model and supervised learning classifiers, significant strides have been made in directly relating and filtering spam dispatches.

The experimental results illuminate the superiority of the logistic regression algorithm in classifying emails into ham or spam orders, showcasing the effectiveness of the proposed approach. also, the study advocates for the wide handover of the BERT model and classifiers in spam discovery due to their remarkable performance.

Likewise, the study opens avenues for future disquisition, including extending the operation of the proposed model to other disciplines analogous as mobile systems and social media platforms for detecting spam dispatches and fake news. also, there is a call for exploring further comprehensive layers within the BERT model to further enhance its effectiveness in text interpretation and point birth.

Overall, this disquisition contributes to the ongoing sweats in combating spam and underscores the eventuality of advanced machine knowledge ways in addressing contemporary challenges in information security and Communication.

REFERENCES

[1] Christina, V., S. Karpagavalli, G. Suganya. "Email spam filtering using supervised machine learning techniques." International Journal on Computer Science and Engineering (IJCSE) 2.09 (2010).

[2] Amandeep Singh Rajput, Vijay Athavale, Sumit Mittal. "Intelligent Model for Classification of SPAM and HAM." IJITEE. ISSN: 2278-3075, Volume-8 Issue-6S, April 2019.

[3] Kumar, N., & Sonowal, S. "Email spam detection using machine learning algorithms." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (2020).

[4] Junnarkar, A., Adhikari, S., Fagania, J., Chimurkar, P., & Karia, D. "E-mail spam classification via machine learning and natural language processing." 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (2021, February).

[5] Awad, W.A., ELseuofi, S.M. "Machine learning methods for spam e-mail classification." Int. J. Comput. Sci. Inf. Technol. (IJCSIT) 3(1), 173–184 (2011).

[6] Zhang, F., Chan, P.P., Biggio, B., Yeung, D.S., Roli, F. "Adversarial feature selection against evasion attacks." IEEE Trans. Cybern.46(3), 766–777 (2015).

[7] Shaukat, K., Luo, S., Chen, S., & Liu, D. "Cyber threat detection using machine learning techniques: A performance evaluation perspective." 2020 international conference on cyber warfare and security (ICCWS) (2020, October).

[8] Garavand, A., Salehnasab, C., Behmanesh, A., Aslani, N., Zadeh, A.H., Ghaderzadeh, M. "Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms." J. Healthc. Eng. (2022).

[9] Ghaderzadeh,M., Aria,M., Asadi, F. "X-ray equipped with artificial intelligence: changing the COVID-19 diagnostic paradigm during the pandemic." BioMed Res. Int. (2021). Hajek, P., Barushka, A., Munk, M. "Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining." Neural Comput. Appl. 32, 17259–17274 (2020).

[10] Ramanathan, V., Wechsler, H. "Phishing detection and impersonated entity discovery using conditional random field and latent Dirichlet allocation." Comput. Secur. 34, 123–139 (2013).

[11] Ghourabi, A., Mahmood, M.A., Alzubi, Q.M. "A hybrid CNNLSTM model for SMS spam detection in Arabic and English messages." Future Internet 12(9), 156 (2020).

[12] Suborna, A.K., Saha, S., Roy, C., Sarkar, S., & Siddique,

[13] M.T.H. "An approach to improve the accuracy of detecting spam in online reviews." 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (2021, February). Frías-Blanco, I., Verdecia-Cabrera, A., Ortiz Díaz, A., & Carvalho, A. "Fast adaptive stacking of ensembles." Proceedings of the 31st Annual ACM Symposium on Applied Computing (2016, April).

[14] El-Kareem, A., Elshenawy, A., Elrfaey, F. "Mail spam detection using stacking classification." J. Al-Azhar Univ. Eng. Sector 12(45), 1242–1255 (2017).

[15] Madichetty, S. "A stacked convolutional neural network for detecting the resource tweets during a disaster." Multimed. Tools Appl. 80, 3927–3949 (2021).

[16] Oh, H. "A YouTube spam comments detection scheme using cascaded ensemble machine

learning model." IEEE Access 9, 144121–144128 (2021).

[17] Zhao, C., Xin, Y., Li, X., Yang, Y., Chen, Y. "A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data." Appl. Sci. 10(3), 936 (2020).

[18] Liu, S., Wang, Y., Zhang, J., Chen, C., Xiang, Y. "Addressing the class imbalance problem in Twitter spam detection using ensemble learning." Comput. Secur. 69, 35–49 (2017).