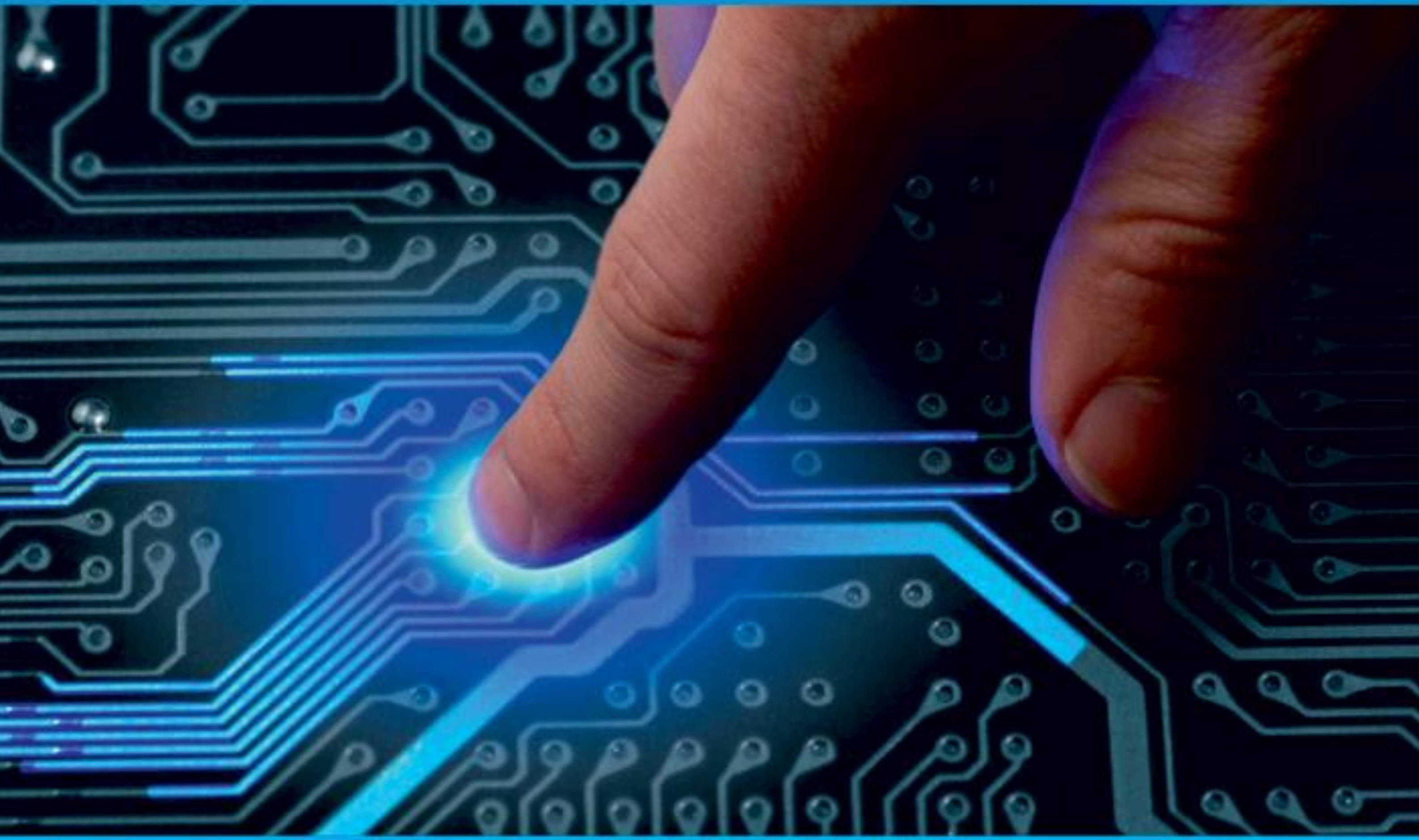
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

A Cross-Lingual Spam Review Detection Approach Using Support Vector Machines

Dr. S. Lavanya, Suriyanaath G, Santhosh V, Selvakumar P

Professor, Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

ABSTRACT: Social media is an effective informational channel for sharing details about the goods and services offered by online retailers. Customers who have purchased the goods themselves offer this information. Analysis of customer-cited features and specifications based on their sentiment. These descriptions and reviews may be found on the Flipkart and Twitter websites. Customers increasingly rely on reviews for product information. However, the usefulness of online reviews is impeded by fake reviews that give an untruthful picture of product quality. Therefore, detection of fake reviews is needed. In today's generation this way of encouraging the consumers to write a review about a product has become a good strategy for marketing their product through real audience's voice. Such precious information has been spammed and manipulated. Out of many researches one fascinating research was done to identify the deceptive opinion spam. Reviews of features/specifications from the Twitter and Flipkart websites were considered for this study project. As a result, the work&#39;s analysis of customers&#39; issues with purchasing high-quality goods was its focus. For the purpose of evaluating comments, this work automates the process of extracting emantic based elements or features and their opinions.

KEYWORDS: Sentiment Analysis, Aspect, Fuzzy Logic, Ecommerce, Customer Reviews, Decision Making

I. INTRODUCTION

In the era of information technology, information sharing has become very easy and fast. Many platforms are available for users to share information anywhere across the world. Among all information sharing mediums, email is the simplest, cheapest, and the most rapid method of information sharing worldwide. But, due to their simplicity, emails are vulnerable to different kinds of attacks, and the most common and dangerous one is spam [1]. No one wants to receive emails not related to their interest because they waste receivers' time and resources. Besides, these emails can have malicious content hidden in the form of attachments or URLs that may lead to the host system's security breaches [2]. Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing [3]. So, it requires an immense demand for the security of the email system. Spam emails may carry viruses, rats, and Trojans. Attackers mostly use this technique for luring users towards online services. They may send spam emails that contain attachments with the multiple-file extension, packed URLs that lead the user to malicious and spamming websites and end up with some sort of data or financial fraud and identify theft [4, 5]. Many email providers allow their users to make keywords base rules that automatically filter emails. Still, this approach is not very useful because it is difficult, and users do not want to customize their emails, due to which spammers attack their email accounts.

In the last few decades, Internet of things (IoT) has become a part of modern life and is growing rapidly. IoT has become an essential component of smart cities. There are a lot of IoT-based social media platforms and applications. Due to the emergence of IoT, spamming problems are increasing at a high rate. The researchers proposed various spam detection methods to detect and filter spam and spammers. Mainly, the existing spam detection methods are divided into two types: behaviour pattern-based approaches and semantic pattern-based approaches. These approaches have their limitations and drawbacks. There has been significant growth in spam emails, along with the rise of the Internet and communication around the globe [6]. Spams are generated from any location of the world with the Internet's help by hiding the attacker's identity. There are a plenty of antispam tools and techniques, but the spam rate is still very high. The most dangerous spams are malicious emails containing links to malicious websites that can harm the victim's data. Spam emails can also slow down the server response by filling up the memory or capacity of servers. To accurately detect spam emails and avoid the rising email spam issues, every organization carefully evaluates the available tools to tackle spam in their environment. Some famous mechanisms to identify and analyze the incoming emails for spam detection are Whitelist/Blacklist [7], mail header analysis, keyword checking, etc.

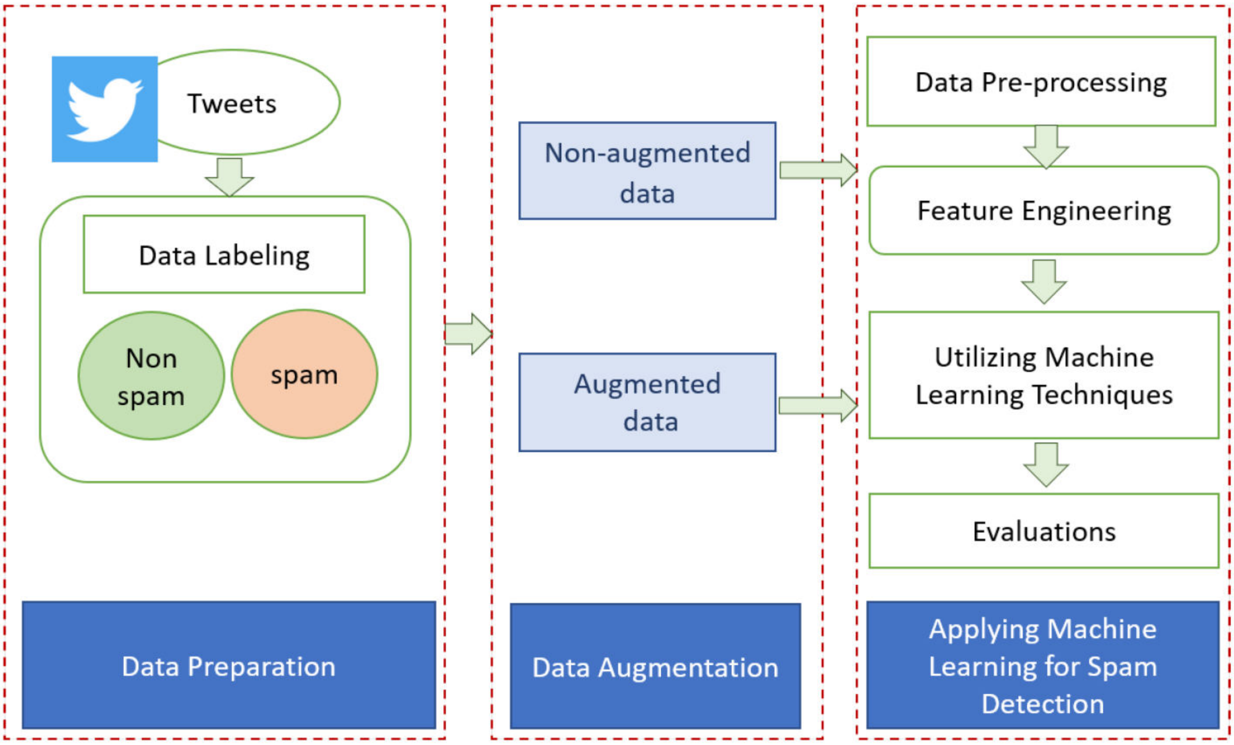


Fig 1: Architecture of the spam detection

Social networking experts estimate that 40% of social network accounts are used for spam [8]. The spammers use popular social networking tools to target specific segments, review pages, or fan pages to send hidden links in the text to pornographic or other product sites designed to sell something from fraudulent accounts. The noxious emails that are sent to the same kind of individuals or associations share regular highlights. By investigating these highlights, one can improve the detection of these types of emails. By utilizing artificial ntelligence (AI) [9], we can classify emails into spam and nonspam emails. This solution is possible by using feature extraction from the messages' headers, subject, and body. After extracting this data based on their nature, we can group them into spam or ham. Today, learning-based classifiers [10] are commonly used for spam detection. In learning-based classification, the detection process assumes that spam emails have a specific set of features that differentiate them from legitimate emails [11]. Many factors increase the complexity of the identification process of spam in learning-based models. These factors include spam subjectivity, idea drift, language problems, overhead processing, and text latency.

II. RELATED WORK

Email spam is nothing more than fake or unwanted bulk mails sent via any account or an automated system. Spam emails are increasing day by day, and it has become a common problem over the last decade. Email IDs receiving spam emails are typically collected through spambots (a computerized application that crawls email addresses across the Internet). The applications of machine learning have been playing a vital role in the detection of spam emails. It has various models and techniques that researchers are using to develop novel spam detection and filtering models [13]. Kaur and Verma [4] present a survey on email spam detection using a supervised approach with feature selection. hey discuss the knowledge discovery process for spam detection systems. They also elaborate various techniques and tools proposed for spam detection. The choice of features based on N-Gram is also addressed in this survey. N-Gram [5, 6] is a predictive-based algorithm used to predict the probability of the next word occurrence after finding $N-1$ terms in a sentence or text corpus. N-Gram uses probability-based techniques for the next word prediction. They compare various machine learning (multilayer perceptron neural network support vector machine, Naïve Bayes) and nonmachine learning (Signatures, Blacklist and Whitelist, and mail header checking) approaches for email spam detection.

Saleh et al. [7] present a survey on intelligent spam email detection. They discuss various security risks of emails, especially spam emails, the scope of spam analysis, and different machine learning and nonmachine learning techniques for spam detection and filtering. They conclude that there is high adoption of supervised learning [18] algorithms for email spam detection. They state that the high usage of supervised learning is the accuracy and consistency of supervised techniques. They also discussed multialgorithm frameworks and found that multialgorithm frameworks are more efficient than a single algorithm. They found that nearly all research work that uses the content of emails for the identification spam, particularly phishing emails, depends on word-based classification or clustering systems.

Blanzieri and Bryl [2, 9] describe a list of learning-based email spam filtering approaches. In this paper, they addressed the spam problems and provided a review of learning-based spam filtering. They explain various features of spam emails. In this study, effects of spam emails on different domains were discussed. Various economic and ethical issues of spam are also discussed in this study. The antispam approach that is common and learning-based filtering is well developed. The commonly used filters are based on different classification techniques applied to various components of email messages. This study suggests that the Naïve Bayes classifier holds a particular position amongst multiple learning algorithms used for spam filtering. With splendid pace and simplicity, it gives high precision results.

Bhuiyan et al. [12] present a review of current email spam filtering approaches. They summarize multiple spam filtering approaches and sum up the accuracy on various parameters of different proposed systems by analyzing numerous processes. They discuss that all the existing methods are efficient for filtering spam emails. Some have successful results, and others are attempting to incorporate other ways to boost their accuracy performance. Although they are all successful, they still have some issues in spam filtering methods, which is the primary concern for researchers. They are trying to create a next-generation spam filtering mechanism to understand large numbers of multimedia data and filter spam emails. They conclude that most email spam filtering is done by utilizing Naïve Bayes and the SVM algorithm. To test the spam filtration models, these models can be trained on different datasets, such as "ECML" and UCI dataset [21].

Ferrag et al. [13] presented a review of deep learning algorithms of intrusion detection systems and spam detection datasets. They discussed various detection systems based on deep learning models and evaluated the effectiveness of those models. They examined 35 well-known cyber dataset by dividing them into seven categories. These categories include Internet traffic-based, network traffic-based, Interanet traffic-based, electrical network-based, virtual private network-based, andriod apps-based, IoT traffic-based, and Internet connected device-based datasets. They conclude that deep learning models can perform better than traditional machine learning and lexicon models for intrusion and spam detection.

III. METHODS

Machine learning is one of the most important and valuable applications of artificial intelligence (AI), which gives computer systems the ability of automatically learning and enhancing their functionality without explicit programming. The primary purpose of machine learning algorithms is to build automated tools to access and use the data for training. The learning process starts with learning labeled data, also called training dataset. It can be a real-life experience, review, example, or feedback to recognize trends in the data to make better future decisions based on the user's input. The main objective of machine learning models is to learn automatically without any intervention from humans. Machine learning consists of three major kinds, used for numerous tasks.

For the last decade, researchers have been trying to make email communication better than today. Spam filtering of emails is one of the most critical ways of protecting email networks. Many research articles have been published using various machine learning approaches to identify and process spam emails, but there are still some research gaps. Junk mail is one of the central, attractive research fields for filling the gaps. For this reason, many spam classification studies have already been carried out using several methods to make email communication more trustworthy and valuable for users. That is why, this paper is presented to make a summarized version of different existing machine learning models and approaches that are being used for email spam detection. This paper also evaluates the most common machine learning approaches like KNN, SVM, random forest, and Naïve Bayes.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204287 |

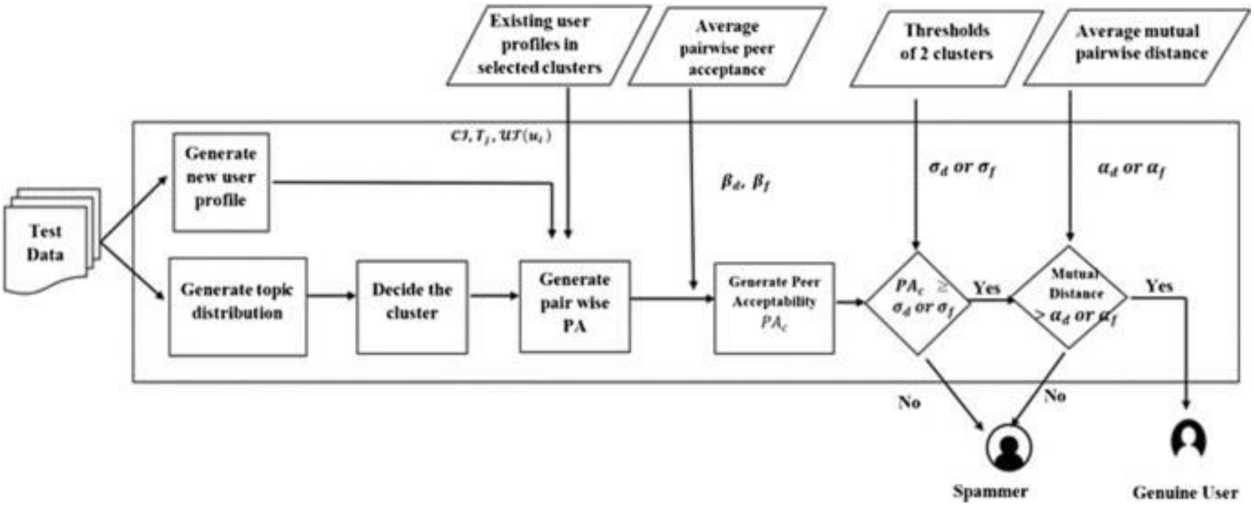

Existing user profiles in selected clusters
Average pairwise peer acceptance
Thresholds of 2 clusters
Average mutual pairwise distance
$CJ, T_j, UT(u_i)$
β_d, β_f
σ_d or σ_f
α_d or α_f
Test Data
Generate new user profile
Generate topic distribution
Decide the cluster
Generate pair wise PA
Generate Peer Acceptability PA_c
$PA_c \geq \sigma_d$ or σ_f
Yes
Mutual Distance $> \alpha_d$ or α_f
Yes
No
No
Spammer
Genuine User

Fig 2: An unsupervised method for social network

Supervised machine learning algorithms are machine learning models that need labeled data. Initially, labeled training data is provided to these models for training, and after training models predict future events. In other words, these models begin with the analysis of an existing training dataset, and they generate a method to make predictions of success values. Upon proper training, the system can provide the prediction on any new data related to the user's data at the training time. Furthermore, the learning algorithm accurately compares the output to the expected output and identifies errors to modify the model.

IV. RESULT ANALYSIS

Developed a spam filtering tool using support vector machine and extreme learning machine algorithms. He used the standard dataset for the development of the spam detection model. SVM got an accuracy of 94.06% in his work, and the extreme learning machine (ELM) model got a 93.04% accuracy level, suggesting just 1.1% performance improvement that SVM achieved over ELM. He indicated that SVM's improvement over ELM accuracy is marginal. It implies that, in situations where detection time is critical, as in real-time systems, the ELM spam detector should be given preference over SVM spam detection. Although SVM got a higher accuracy level in his research, it takes more time for training than the ELM system. Tretyakov [62] also discussed various machine learning techniques for email spam filtering. This paper compared the precision results between false positives and precision results after eliminating false positives. They show the result after eliminating false positives, which were more accurate and reliable than before.

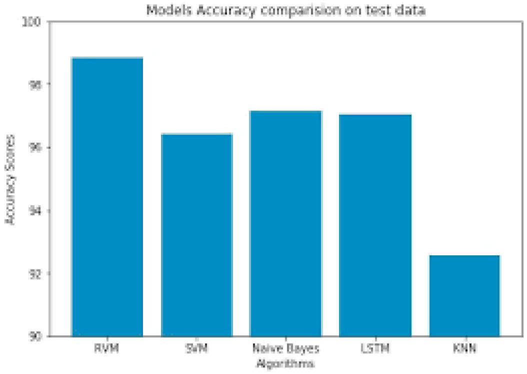


Fig 3: Result analysis of Spam Detection using Relevance

Presented an article on machine learning-based spam detection techniques for IoT devices. They used five ML models and analyzed their results using various performance metrics. A large number of input features were used for the training of proposed models. Each model calculates a spam score based on the input attributes. This score represents the trustworthiness of an IoT device based on a variety of factors. The suggested approach is validated using the REFIT smart home dataset. They claim that their proposed system can detect spam better than currently used spam detection systems. Their work can be utilized in smart homes and other places where intelligent devices are used.

IV. CONCLUSION

This section discusses the research gaps and open research problems of the spam detection and filtration domain. In the future, experiments and models should be trained on real-life data rather than manually created datasets, because, in the various article, the models trained on artificial datasets perform very poorly on real-life data. Currently, supervised, unsupervised, and reinforcement learning algorithms are used for spam detection, but we can get higher accuracy and efficiency by using hybrid algorithms in the future. Feature extraction can be improved in the future by using deep learning for feature extraction. Using clustering techniques for spam filtering relevance feedback using dynamic updating can better cluster spam and ham. Along with machine learning, blockchain models and concepts can also be used for email spam detection in the future. Experts in linguistics and psycholinguistics can collaborate in the future for manual annotation of datasets, which will result in the development of effective and standard spam datasets with high dimensionality. In future, spam filters can be designed with faster processing and classification accuracy using Graphics Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs).

REFERENCES

1. H. Faris, A. M. Al-Zoubi, A. A. Heidari et al., “An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks,” *Information Fusion*, vol. 48, pp. 67–83, 2019.
   View at: Publisher Site | Google Scholar
2. E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.
   View at: Publisher Site | Google Scholar
3. A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, “Email classification using artificial neural network,” *International Journal for Academic Development*, vol. 2, 2018.
   View at: Google Scholar
4. N. Udayakumar, S. Anandaselvi, and T. Subbulakshmi, “Dynamic malware analysis using machine learning algorithm,” in *Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, Palladam, India, December 2017.
   View at: Google Scholar
5. S. O. Olatunji, “Extreme Learning machines and Support Vector Machines models for email spam detection,” in *Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, IEEE, Windsor, Canada, April 2017.
   View at: Google Scholar
6. J. Dean, “Large scale deep learning,” in *Proceedings of the Keynote GPU Technical Conference*, San Jose, CA, USA, 2015.
   View at: Google Scholar
7. J. K. Kruschke and T. M. Liddell, “Bayesian data analysis for newcomers,” *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 155–177, 2018.
   View at: Publisher Site | Google Scholar
8. K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, “Malicious accounts: dark of the social networks,” *Journal of Network and Computer Applications*, vol. 79, pp. 41–67, 2017.
   View at: Publisher Site | Google Scholar
9. A. Barushka and P. Hájek, “Spam filtering using regularized neural networks with rectified linear units,” in *Proceedings of the Conference of the Italian Association for Artificial Intelligence*, Springer, Berlin, Germany, November 2016.
   View at: Google Scholar
10. F. Jamil, H. K. Kahng, S. Kim, and D. H. Kim, “Towards secure fitness framework based on IoT-enabled blockchain network integrated with machine learning algorithms,” *Sensors*, vol. 21, no. 5, p. 1640, 2021.
    View at: Google Scholar

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details