Improving Email Spam Prediction with Novel Recurrent Neural Network versus Logistic Regression

Chillakuru Neeharika and S. Kalaiarasi
Saveetha University, India

Keywords: Electronic Mail, Logistic Regression, Vulnerability, Recurrent Neural Network, Machine Learning, Spam, Unsupervised Approach.

Abstract: The goal of the research is to improve accuracy in detection of email spam using a novel recurrent neural network compared to logistic regression. Two groups such as Novel Recurrent Neural Network and Logistic regression are taken for this study. The sample size for each group is 10, and the study's parameters include an alpha value of 0.8 and a beta value of 0.2. About 80% is the G-power value. In terms of predicting spam in electronic mail, the Novel Recurrent Neural Network has the highest accuracy (97.36%), while the Logistic Regression comes in at 94.90%. $p=0.005(p0.05)$, which is the statistically significant value. The Novel Recurrent Neural Network better than the Logistic Regression in Electronic mail spam prediction.

1 INTRODUCTION

Spam email is unsolicited and undesired junk email sent in large quantities or in bulk to an indiscriminate recipient list (Dhinakaran et al. 2007). The adversary intentionally modifies the data to deceive the classifiers by taking advantage of the dataset shift vulnerability. Spam is transmitted for monetary gain. Botnets, or networks of compromised machines, send it in large quantities (Rayan 2022). Spam email is frequently a vulnerable effort to obtain unauthorized access to your system (Weiske et al. 2020). Spam prohibits users from making full and effective use of their CPU time, storage capacity, and network bandwidth(Jeong 2012). It becomes a major issue, especially when spam communications are mixed in with crucial business emails (Sroufe et al. 2009). Dealing with difficulties caused by spam email becomes unavoidable(Cota and Zinca 2022). Spammers use the web security flaw known as the "email vulnerability" to send out anonymous emails. As a result, this problem may be handled by employing Machine Learning approaches capable of detecting and filtering spam (Sultana, Yenepoya Institute of Technology, and Moodbidri 2020) (G. Ramkumar et al 2022). Recent research has demonstrated the vulnerability of machine learning algorithms to adversarial attacks, wherein minor input perturbations result in misclassification. There are a variety of classes for vulnerabilities where the attacks are found at the detection phase (AbdulNabi and Yaseen 2021) (Padma, S et al. 2022).

In 765 scholarly articles, same method is used to enhance email spammer prediction. There are various algorithms for predicting email spam, but a few of them are the Novel Recurrent Neural Network and the supervised and unsupervised approaches to logistic regression (Honan and Curran 2009). Novel Recurrent Neural Network (RNN) is a method that is unsupervised. It performs a variety of tasks, including storing data while input is read incrementally and generalizing model services to input configurations (Williams et al. 2019). The primary purpose of a novel RNN is to replace unidirectional networks by allowing data to flow from one layer to another (Rajalingam 2020). The other algorithm is supervised and uses logistic regression. The association between the categorical dependent variable and the dependent variable is measured by logistic regression (Butt et al. Messages were marked and are uncertain about new incoming messages for the Spam detection issue. In this sense, a model that can determine if a communication is spam or not is necessary. The logistic regression model is employed for this, with 0 denoting negative class that is spam message is not present and 1 denoting positive class indicating that the spam is present. (Chen and Yang 2022).

The biggest drawback is that due to their propensity for overfitting, neural networks may be overly impacted by their training data and hence struggle to generalize to new, untried data. A higher

Neeharika, C. and Kalaiarasi, S.
Improving Email Spam Prediction with Novel Recurrent Neural Network versus Logistic Regression.
DOI: 10.5220/0012566500003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 91-95
ISBN: 978-989-758-661-3

rate of false positives or false negatives may come from this. So, it seeks to use machine learning to increase the suggested system's accuracy. The aim is to use Novel Recurrent Neural Networks to predict Email Spam than Logistic Regression more accurately.

2 MATERIALS AND METHODS

Novel Recurrent Neural Network

Novel Recurrent Neural Network can act as a Feedforward neural networks offspring. It may handle variable length input sequences by employing internal state memory (Khurana et al. 2022). As a result, they can be used for tasks like unsegmented, connected handwriting recognition and speech recognition (Maleh et al. 2020). The disadvantage of Recurrent Neural Network is it can be slow due to the sequential nature of the data. Novel Recurrent Neural Networks are turing complete in theory and can execute arbitrary programmes to process arbitrary input sequences. It refers to the class of networks with an infinite impulse response (Sarno et al. 2020).

Algorithm

Step 1. Import the required libraries and read the datasets of the plant images.

Step 2. Load and split the dataset into training and testing of the plant disease detection.

Step 3. The image will be preprocessed.

Step 4. Prepare the input image to the required Keras format to perform features and transform the data.

Step 5. Create and initialize the RNN model and train it with the image of the leaf.

Step 6. Make the prediction on the training and testing of the datasets.

Step 7. The model’s precision score will be examined and resulted in the trained graph with matplotlib.

Logistic Regression

Based on a collection of independent variables, it is used to forecast the categorical dependent variable (O'Neil and Schutt 2013). Using logistic regression, the result of a categorical dependent variable is predicted. Therefore, the output must be discrete or categorical. It provides the probabilistic values that fall between 0 and 1 (Teja and Sai Teja 2021). It can be Yes or No, 0 or 1, true or false, and so on. The method of application is the only distinction between linear regression and logistic regression. Classification issues are resolved using logistic regression, and regression issues are resolved using linear regression. The disadvantage of Logistic regression can be affected by irrelevant features, leading to a decrease in accuracy (Rajendran et al. 2016).

Logistic Regression Algorithm Steps

Step 1. Design the LR Model with Predefined libraries.

Step 2. Read Train Data set containing heart disease related to past year data.

Step 3. Obtain the heart disease rate.

Step 4. Train the model with data set and compute the prediction factor values

Step 5. Assign test data to the model and perform classification

Step 6. Compare the similarities between test and train data

Step 7. Write the results of prediction values.

Table 1: For novel recurrent neural networks and logistic regression algorithms, the iteration values.

Iterations	**Novel Recurrent Neural Network**	**Logistic regression**
1	96.20	93.60
2	96.20	93.60
3	96.20	93.60
4	96.10	93.60
5	99.30	94.50
6	96.20	95.00
7	99.10	95.30
8	99.10	95.80
9	99.20	96.60
10	99.30	97.40

3 RESULTS

The statistical results obtained from this study can be applied to all variables. Based on the outcomes of the independent sample t-test and Logistic regression analysis, it is evident that the enhanced Novel RNN model has achieved the better accuracy and standard deviation compared to the Logistic regression model. The difference in accuracy between the two models can be attributed to the significant and interconnected nature of the study outcomes, which highlights the importance of adhering to the equality of variance principle in such analyses.

Table 2: Comparison of the Novel Recurrent Neural Network (97%) method with Logistic Regression (94%) for 10 iterations using group statistics.

	Group Name	N	Mean	Standard Deviation	Standard Error Mean
Accuracy	Novel Recurrent Neural Network	10	97.96	1.54	0.48
	Logistic Regression	10	94.90	1.37	0.43

Table 1 displays the results of the study, indicating that Novel RNN has a higher accuracy rate of 97.96%, outperforming Logistic Regression model, which has an accuracy of 94.90%. This highlights the effectiveness of Novel RNN in categorizing data compared to the Logistic Regression model. Moreover, Table 2 shows that the Standard Deviation of the Novel Recurrent Neural Network is 1.55599, while the Logistic Regression model has a Standard Deviation of 1.86118. The independent samples T-test was utilized to compare the performance of the two models, where the Mean difference was 2.511, and the standard error difference was 0.77856. In Table 3, However, the significance value was found to be p=0.005 (p<0.05), implying that there is a significant difference between the two models. Figure 1 infers the comparison of accuracy of Novel RNN and Logistic Regression.

Table 3: T-test results from independent samples reveal statistically significant value 0.005(p≤0.05), the mean difference is 3.06000, and the difference in standard error is 0.65323.

Independent Sample Test										
Levene's Test for Equality of Variances				T-test for Equality of Means						
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Differences	95% Confidence Interval of the Difference	
									Lower	Upper
Accuracy	Equal Variances assumed	0.002	0.962	3.225	18	0.005	2.51111	0.77856	0.87541	4.14681
	Equal Variances not assumed			3.286	17.981	0.005	2.51111	0.76415	0.90557	4.11665

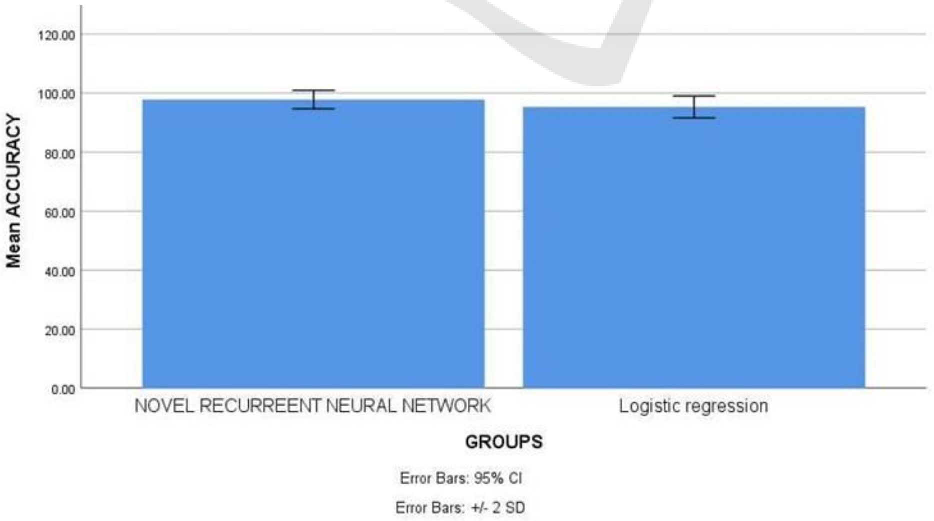

Figure 1: Novel Recurrent Neural Network and Logistic Regression are represented on the X-axis. The mean accuracy of the novel recurrent neural network and logistic regression is displayed on the Y-axis, Mean Efficiency of detection is ±2 SD.

4 DISCUSSION

The study conducted to predict email spam using the Novel Recurrent Neural Network showed that the unsupervised approach had significantly higher accuracy compared to the supervised approach, with an accuracy rate of around 97% versus 94% for Logistic regression. However, it should be noted that Logistic regression has limitations in achieving high accuracy rates. On the other hand, the Novel Recurrent Neural Network (RNN) tends to provide more consistent outcomes, as evidenced by its lower standard deviation (Broadhurst and Trivedi 2020).

The results of the study indicated that Novel RNN achieved an accuracy rate of 97% for Email spam prediction, which is equivalent to the findings presented in the paper. In contrast, the reported Logistic regression model had an accuracy rate of 94% for the same task of Email spam prediction. The RNN, Logistic regression is a parameter used to predict Email spam (Wang and Katagishi 2014). Using Logistic regression for Email spam prediction will have significant concerns to pretend that this innovation reveals that logistic regression has the least accuracy of 94%.

The disadvantage of Logistic regression is that increasing the value of the dataset only tends to achieve the necessary precision. Novel Recurrent Neural Network works better when combined with other techniques (Kigerl 2018). Irrelevant features can degrade the accuracy of logistic regression (Rafat et al. 2022). Our future innovation will concentrate on improving accuracy for predicting Email spam without any disadvantages in working mode (Kaddoura et al. 2022).

5 CONCLUSION

Finding out how successfully Novel Recurrent Neural Network and Logistic Regression, ave predicted email spam is the goal of the current study. The highest accuracy of 97.96% was provided by the RNN, compared to the Logistic regression accuracy of 94.90%. This shows that the performance of predicting the email spam is good for the Novel RNN.

REFERENCES

AbdulNabi, Isra ’a, and Qussai Yaseen. (2021). “Spam Email Detection Using Deep Learning Techniques.” Procedia Computer Science. https://doi.org/10.1016/j.procs.2021.03.107.

Broadhurst, Roderic, and Harshit Trivedi. (2020). “Malware in Spam Email: Risks and Trends in the Australian Spam Intelligence Database.” https://doi.org/10.52922/ti04657.

Butt, Umer Ahmed, Rashid Amin, Hamza Aldabbas, Senthilkumar Mohan, Bader Alouffi, and Ali Ahmadian. (2022). “Cloud-Based Email Phishing Attack Using Machine and Deep Learning Algorithm.” Complex & Intelligent Systems, June, 1–28.

Chen, Yanfang, and Yongzhao Yang. (2022). “An Advanced Deep Attention Collaborative Mechanism for Secure Educational Email Services.” Computational Intelligence and Neuroscience 2022 (April): 3150626.

Cormack, Gordon V. (2008). Email Spam Filtering: A Systematic Review. Now Publishers Inc.

Cota, Rodica Paula, and Daniel Zinca. 2022. “Comparative Results of Spam Email Detection Using Machine Learning Algorithms.” 2022 14th International Conference on Communications (COMM). https://doi.org/10.1109/comm54429.2022.9817305.

Dhavale, and Sunita Vikrant. (2017). Advanced Image-Based Spam Detection and Filtering Techniques. IGI Global.

Dhinakaran, Cynthia, Cheol-Joo Chae, Jae-Kwang Lee, and Dhinaharan Nagamalai. (2007). “An Empirical Study of Spam and Spam Vulnerable Email Accounts.” Future Generation Communication and Networking (FGCN 2007). https://doi.org/10.1109/fgcn.2007.61.

G. Ramkumar, G. Anitha, P. Nirmala, S. Ramesh and M. Tamilselvi, (2022)"An Effective Copyright Management Principle using Intelligent Wavelet Transformation based Water marking Scheme," International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2022, pp. 1-7, doi: 10.1109/ACCAI53970.2022.9752516.

Honan, John, and Kevin Curran. (2009). “The Problem of Spam Email.” Understanding the Internet. https://doi.org/10.1016/b978-1-84334-499-5.50022-2.

Jeong, Hyuncheol. (2012). “Detection of Zombie PCs Based on Email Spam Analysis.” KSII Transactions on Internet and Information Systems. https://doi.org/10.3837/tiis.2012.05.011.

Kaddoura, Sanaa, Ganesh Chandrasekaran, Daniela Elena Popescu, and Jude Hemanth Duraisamy. 2022. “A Systematic Literature Review on Spam Content Detection and Classification.” PeerJ. Computer Science 8 (January): e830.

Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. “Natural Language Processing: State of the Art, Current Trends and Challenges.” Multimedia Tools and Applications, July, 1–32.

Kigerl, Alex C. 2018. “Email Spam Origins: Does the CAN SPAM Act Shift Spam beyond United States Jurisdiction?” Trends in Organized Crime. https://doi.org/10.1007/s12117-016-9289-9.

Listík, Vít, Jan Šedivý, and Václav Hlaváč. 2020. “Email Image Spam Classification Based on ResNet

Convolutional Neural Network." Proceedings of the 6th International Conference on Information Systems Security and Privacy. https://doi.org/10.5220/0008956704570464.

Maleh, Yassine, Mohammad Shojafar, Mamoun Alazab, and Youssef Baddi. 2020. Machine Intelligence and Big Data Analytics for Cybersecurity Applications. Springer Nature.

O'Neil, Cathy, and Rachel Schutt. 2013. Doing Data Science: Straight Talk from the Frontline. "O'Reilly Media, Inc."

Padma, S., Vidhya Lakshmi, S., Prakash, R., Srividhya, S., Sivakumar, A. A., Divyah, N., ... & Saavedra Flores, E. I. (2022). Simulation of land use/land cover dynamics using Google Earth data and QGIS: a case study on outer ring road, Southern India. Sustainability, 14(24), 16373

"PERFORMANCE OF MACHINE LEARNING TECHNIQUES FOR EMAIL SPAM FILTERING." 2018. International Journal of Recent Trends in Engineering and Research. https://doi.org/10.23883/ijrter.conf.20171201.049.yzvdv.

Rafat, Khan Farhan, Qin Xin, Abdul Rehman Javed, Zunera Jalil, and Rana Zeeshan Ahmad. 2022. "Evading Obscure Communication from Spam Emails." Mathematical Biosciences and Engineering: MBE 19 (2): 1926–43.

Rajalingam, Mallika. 2020. Text Segmentation and Recognition for Enhanced Image Spam Detection: An Integrated Approach. Springer Nature.

Rajendran, P., M. Janaki, S. M. Hemalatha, and B. Durkananthini. 2016. "Adaptive Privacy Policy Prediction for Email Spam Filtering." 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave). https://doi.org/10.1109/startup.2016.7583948.

Rayan, Alanazi. 2022. "Analysis of E-Mail Spam Detection Using a Novel Machine Learning-Based Hybrid Bagging Technique." Computational Intelligence and Neuroscience 2022 (August): 2500772.

Sarno, Dawn M., Joanna E. Lewis, Corey J. Bohil, and Mark B. Neider. 2020. "Which Phish Is on the Hook? Phishing Vulnerability for Older Versus Younger Adults." Human Factors 62 (5): 704–17.

Sroufe, Paul, Santi Phithakkitnukoon, Ram Dantu, and Joao Cangussu. (2009). "Email Shape Analysis for Spam Botnet Detection." 2009 6th IEEE Consumer Communications and Networking Conference. https://doi.org/10.1109/ccnc.2009.4784781.

Sultana, Thashina, Yenepoya Institute of Technology, and Moodbidri. (2020). "Email Based Spam Detection." International Journal of Engineering Research and. https://doi.org/10.17577/ijertv9is060087.

Teja, P. Sai, and P. Sai Teja. 2021. "Prediction of Spam Email Using Machine Learning Classification Algorithm." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.35226.

United States. Congress. Senate. Committee on Commerce, Science, and Transportation. (2013). Spam (unsolicited Commercial E-Mail): Hearing Before the Committee on Commerce, Science, and Transportation, United States Senate, One Hundred Eighth Congress, First Session, May 21, 2003.

Wang, Jianyi, and Kazuki Katagishi. (2014). "Image Content-Based 'Email Spam Image' Filtering." Journal of Advances in Computer Networks. https://doi.org/10.7763/jacn.2014.v2.92.

Weiske, Rebecca, Maureen Sroufe, Mindy Quigley, Theresa Pancotto, Stephen Werre, and John H. Rossmeisl. (2020). "Development and Evaluation of a Caregiver Reported Quality of Life Assessment Instrument in Dogs with Intracranial Disease." Frontiers in Veterinary Science 7 (August): 537.

Williams, Sarah E., Dawn M. Sarno, Joanna E. Lewis, Mindy K. Shoss, Mark B. Neider, and Corey J. Bohil. (2019). "The Psychological Interaction of Spam Email Features." Ergonomics62 (8): 983–94.