Email spam detection using machine learning algorithms

[1]Jayendar .V, [2]K. Karthik reddy, [3]K.Kuladeep, [4]G.laxma reddy, [5]Mrs.Kalpana Ragutla,
[1,2,3,4] U.G.Scholor, Department of ECE, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.
[5]Research Guide, Department of ECE, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.

ABSTRACT

Our primary objective in this study is to develop a novel approach to detecting spam emails by integrating support vector machines with linear regression. We will determine how successful it is. Procedures and Materials: Both approaches detect spam emails by using linear regression and state-of-the-art support vector machines. There are 10 people in the sample. The predictions made by SPSS seem to be valid, given the high level of confidence (95% CI and 80% G power). With a p-value of p=0.105 (p>0.05) after 10 rounds of the procedure, we were unable to achieve statistical significance. The innovative SVM approach outperformed LR in identifying spam emails, with an accuracy rate of 86.34 percent. In terms of spam email identification, the results showed that the Novel SVM Algorithm performed better than the LR Algorithm.

Keywords Linear regression, spam email, machine learning, new SVM, and exploitation are all synonyms for one other.

INTRODUCTION

Email is among the most popular means of company-to-business communication [1]. What some call "junk email" really refers to is unsolicited mass emailing sent to an anonymous subscriber list. "Junk emails," or unsolicited marketing messages providing various services (such as debt relief, online dating, healthcare product exploitation, etc.), have proliferated with the ubiquity of email. Companies are always looking for new ways to protect their customers' computers against spam emails, which might include viruses [3]. It is not a novel idea to have automated trash recognition. And it achieves all that and more by stopping the waste of time and important network resources. In addition, malicious software, phishing attempts, cross-site scripting, and cross-site request forgery may all be sent via spam email [4]. Inappropriate advertisements and product sales are much more problematic than unsolicited mail, which may include hazardous information, according to data [5]. Numerous sectors may benefit from spam detection systems. The 11,655 scholarly journal articles and conference proceedings published in the last five years formed the basis of our study. Springer, Google Scholar, the ACM Digital Library, IEEE Xplore, and countless more are among the various sources that formed the basis of these works. When it comes to the issue of spam email detection utilizing ML approaches, all three of these authors have provided workable answers. When compared to the current system's pitiful spam detection abilities, the proposed solution is leagues ahead of the competition in terms [6]. The

categorization and implementation of this strategy have lately changed in several ways. One potential application of machine learning is in the identification and filtering of spam. Another objective of this study's dataset classification is the detection of spam emails using LR and SVM algorithms. In terms of traffic and citations, this type of study may provide respectable profits. Multiple studies have investigated intelligent spam email detection. [7]. Future applications of supervised machine learning for spam email detection are detailed in the study. Looking at spammers' techniques for sending unwanted emails, categorizing datasets, and tracking new problems. The identification of fake email messages by the use of distributed word embedding and deep learning [8].

More study is needed to enhance the existing system, since previous studies found that spam email detection was not very accurate. But this doesn't rule out the possibility that Novel Support Vector Machine may remain helpful in the fight against spam emails. [9]. One major advantage is that data classification tasks might potentially provide reliable results by using machine learning techniques. This approach has the potential to improve spam email detection by using Novel Support Vector Machine technology [10].

MATERIALS AND METHODS

The Saveetha School of Engineering's research staff included computer scientists and engineers. The Open Source Lab of the Saveetha Institute is associated with this department. Individual investigations of this matter are underway at a number of research institutions. A comparison was made between the Linear Regression-based Novel Support Vector Machines developed by Group 2 and Group 1. A. Zamir et al. (2020) used a dataset consisting of spam email and ran 20 iterations of Linear Regression using Novel Support Vector Machine technology at different intervals. The calculation was done using the 95% confidence interval, which has a beta of 0.2 and an alpha of 0.05.

Researchers used a real-time dataset that included a large number of spam emails to conduct the research. information retrieved from a CSV file hosted on kaggle.com [11]. "Junk email" significantly increased accuracy compared to the other aspects. Due to post-collection preprocessing of the meteorological information, there is substantial dispute about the attribute definition. Data was vectorized after cleaned-up strings, words, and characters were converted to integers using feature extraction. The machine learning approach was able to operate correctly since the dataset did not include any null or empty values. As we were finalizing the preparations, the dataset was halved. The outcome was that testing only used 20% of the dataset, whereas training used 80%.

Visual Studio Code and Google CoLab were the tools we used for our assessment.We have computers that come with 64-bit OSes already installed. The app will always utilize Windows 10 during installation.

Novel Support Vector Machine (SVM)

Novel SVM is far and away the best supervised machine learning method currently available. Support vector machines (SVMs) mainly aim to assign a numerical value between 0 and 1 to outcomes, which signifies the success or failure of the event. The whole thing is a multidimensional hyperplane that represents several classes when put together. Multiple constructions of the hyperplane will be carried out to guarantee correctness. The discovery of MMH relies on the categorization of datasets. The input data space is transformed into the required format using a linear kernel. Table 1 details one innovative support vector machine algorithm.

Linear Regression (LR)

One way to illustrate the relationship between an explanatory variable or factors and a scalar response is using linear regression. This technique uses a linear approach with both independent and dependent variables. When only one explanatory variable is present, basic linear regression is used. Multiple linear regression is used when there are several explanatory factors. Instead of assuming a single scalar variable, as is the case with multivariate linear regression, this method takes into account many correlation based variables. By analyzing data, linear predictors may attempt to forecast version characteristics that are not yet known, much like linear regression.

Statistical Analysis

We will use IBM SPSS V26.0 to do the statistical analysis. In the social science statistical package, "accuracy" is the dependent variable for mean and other statistical calculations, along with "attachment," "date," and "address" as the independent variables. Each group uses accuracy as the dependent variable and goes through 10 rounds of SPSS to create the dataset [12].

RESULTS

The New SVM Algorithm outperforms LR when it comes to spam email detection.

The new support vector machine algorithm's findings are shown in Table 1. The program builds a database of spam emails and uses it to identify them.

Section 2: The Simple Linear Regression Approach Section 2. The program builds a database of spam emails and uses it to identify them.

Enhanced Email Junk Detection Accuracy (Table 3), with Novel Support Vector Machine achieving 86.43% and Linear Regression 81.67%.

With a p-value of 0.105, the T-test (Table 4) demonstrated that Novel Support Vector Machine outperformed Linear Regression when the dataset's confidence interval was changed to 95%. We utilized a 5% significance threshold since $p>0.05$.

The results of a mixed-methods statistical study using Novel SVM and LR are shown in Table 5. By the conclusion of the tenth cycle, both Linear Regression and Novel SVM have attained accuracy, dispersion, and average standard deviation. Modern support vector machine

outperformed older, more manual linear regression techniques. Figure 1 shows that compared to results obtained using Novel Support Vector Machine, Linear Regression often produces worse results. Over LR, Novel Support Vector Machine Outperforms in Mean and Standard Deviation Comparisons. On the X-axis, the GROUP test displays the results of a comparison between Novel SVM and Linear Regression. With a 95% confidence range of +/- 2 standard deviations, the mean detection accuracy is shown on the Y-axis.

DISCUSSION

In this work, SPSS is used to do the statistical analysis. Novel Support Vector Machine is used for context determination. Garbage sorting is accomplished using Linear Regression (LR). Methods for employing machine learning to categorize emails as spam or garbage have been the primary focus of most academic study on the problem. With an astounding 86.34% accuracy rate on the dataset, the Novel Support Vector Machine approach for email recognition well surpasses Linear Regression's 81.67% performance. Because both the SVM and LR approaches provide the same statistical result ($p>0.05$), we can say that they are statistically equivalent. Their team is working on a new Support Vector Machine (SVM) algorithm with the expectation that it would improve spam and unwanted email detection. For the obvious reason that data categorization jobs are a perfect fit for support vector machines (SVMs), which can learn complex decision boundaries. [13]. When dealing with previously classified sensitive material, their expertise is invaluable. Experimental findings demonstrated that the Novel SVM approach achieved better accuracy, precision, recall, and F1 score than the linear regression strategy on a publicly accessible spam dataset. By projecting the data onto a higher-dimensional space, kernel functions may be useful in cases when the data isn't linearly separable [14]. Whereas, the linear regression model allows for the prediction of continuous outcomes using a collection of variables. Problems with task classification resulting from non-linearly separable data cannot be solved using linear regression because it is only able to learn decision boundaries. Ignoring spam emails with intricate patterns that a basic model would miss is a real possibility because of this. [15].

The research has one little flaw: it uses the provided context to classify and assess spam email detection. The most common problems include failing to install the system and struggling to categorize complex data that SVM does not understand. [16].Because this makes data identification, distribution, and classification easier, it will be helpful for future research.

CONCLUSION

Using Novel SVM and LR, we aim to improve the accuracy of spam email detection in this work. When tested on the given dataset, the Novel SVM technique outperformed the LR method in terms of detection accuracy by a margin of 86.34%.

REFERENCES

[1] M. Ramprasad, N. Chowdary, K. Reddy, and V. Gaurav, “EMAIL SPAM DETECTION

USING PYTHON & MACHINE LEARNING," 2021, Accessed: Dec. 26, 2022. [Online]. Available: https://www.semanticscholar.org/paper/EMAIL-SPAM-DETECTION-USING-PYTHON-%26-MACHINE-Ramprasad-Chowdary/07e029dc33c72c0ef206d34b9377ee417e938392

[2] N. Oswal and Y. Sharma, "Email Spam Detection and Filtering Using Machine Learning," 2021, Accessed: Dec. 26, 2022. [Online]. Available: https://www.semanticscholar.org/paper/Email-Spam-Detection-and-Filtering-Using-Machine-Oswal-Sharma/320175e49169d0d6ebe11afe8b23169888e9c772

[3] S. S. Roy, A. Sinha, R. Roy, C. Barna, and P. Samui, "Spam Email Detection Using Deep Support Vector Machine, Support Vector Machine and Artificial Neural Network," 2016, doi: 10.1007/978-3-319-62524-9_13.

[4] R. Nayak, S. Jiwani, and B. Rajitha, "Spam email detection using machine learning algorithm," *Materials Today: Proceedings*, 2021, doi: 10.1016/J.MATPR.2021.03.147.

[5] M. Mohammed, D. Ibrahim, and A. O. Salman, "Adaptive intelligent learning approach based on visual anti-spam email model for multi-natural language," *Journal of Intelligent Systems*, 2021, doi: 10.1515/jisys-2021-0045.

[6] I. AbdulNabi and Q. M. Yaseen, "Spam Email Detection Using Deep Learning Techniques," 2021, doi: 10.1016/j.procs.2021.03.107.

[7] A. Tabish, "Machine Learning Techniques for Spam Detection in Email," *Medium*, Aug. 23, 2022. https://medium.com/@alinatabish/machine-learning-techniques-for-spam-detection-in-email-7db87eb11bc2 (accessed Dec. 22, 2022).

[8] S. Srinivasan, V. Ravi, M. Alazab, S. Ketha, A. M. Al-Zoubi, and K. Padannayil, "Spam Emails Detection Based on Distributed Word Embedding with Deep Learning," *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*, pp. 161–189, 2021.

[9] A. Zamir, H. U. Khan, W. Mehmood, T. Iqbal, and A. U. Akram, "A feature-centric spam email detection model using diverse supervised machine learning algorithms," *The Electronic Library*, vol. 38, no. 3, pp. 633–657, Jul. 2020.

[10] S. V. S. Bharath, "Junk Email Detection.csv." Dec. 21, 2022. Accessed: Dec. 23, 2022. [Online]. Available: https://www.kaggle.com/bharathsaivs/junkcsv

[11] D. P. Mood, J. R. Morrow Jr, and M. B. McQueen, *Introduction to Statistics in Human Performance: Using SPSS and R*. Routledge, 2019.

[12] A. Dhatark, S. Pandey, and R. Shinde, "Spam Email Detection Using Machine Learning," *International Journal of Advanced Research in Science, Communication and Technology*, 2022, doi: 10.48175/ijarsct-3981.

TABLES & FIGURES

Table 1. Algorithm for Novel SVM. The software collects spam emails and stores them in a database for later use.

Input: Junk Email Detection Dataset
Output: Better Accuracy.
The first stage is to collect an email dataset that contains both legitimate and spam communications. In order to get the emails ready for classification, the second step is to extract the required attributes. Information like the contents of the email, the domain name of the sender, the existence of certain letters or sequences, etc. Part two involves splitting the dataset in half. I give you permission to train and test on your own now. Step four involves constructing a support vector machine classifier using the acquired features. Over the training data, this will be applied. Finally, assess the classifier's efficacy using the test set. Modifying the classifier's features and hyperparameters is the sixth stage. Step 7: Create a trash and non-junk folder for newly received emails using the classifier. Step eight entails checking the precision.

Table 2. A Method for Linear Regression. The program builds a database of spam emails and uses it to identify them.

Input: Junk Email Detection Dataset
Output: Better Accuracy.
The first step need to be to compile an email database that includes spam as well. In order to prepare the emails for classification, the second step is to extract characteristics. This will include every single word of the email, down to the sender's website name and any special characters or strings found therein. Finally, divide the dataset in half. All you have to do is run your operations on either the training set or the test set. Building a linear regression model using the acquired attributes and training data is the fourth stage. Applying the fifth step, evaluate how well the classifier performs. In the sixth phase, you may tweak the classifier's hyperparameters and features to make it work better. In Step 7, we'll turn on the model that checks incoming emails for quality. Determining the precision is the ninth stage.

Table 3. Novel Support Vector Machine achieved 86.43% accuracy and Linear Regression 81.67% accuracy, significantly improving the accuracy of Junk Email Detection.

The Next Version	Level of Accuracy for a New Support Vector Machine (%)	Proportion of Correct Linear Regressions
1	97.30	88.60
2	95.63	88.10
3	92.44	87.00
4	88.60	84.20
5	86.10	81.60

6	84.60	79.60
7	82.70	78.30
8	80.89	77.20
9	78.56	76.30
10	76.58	75.80
Accuracy	86.3400	81.6700

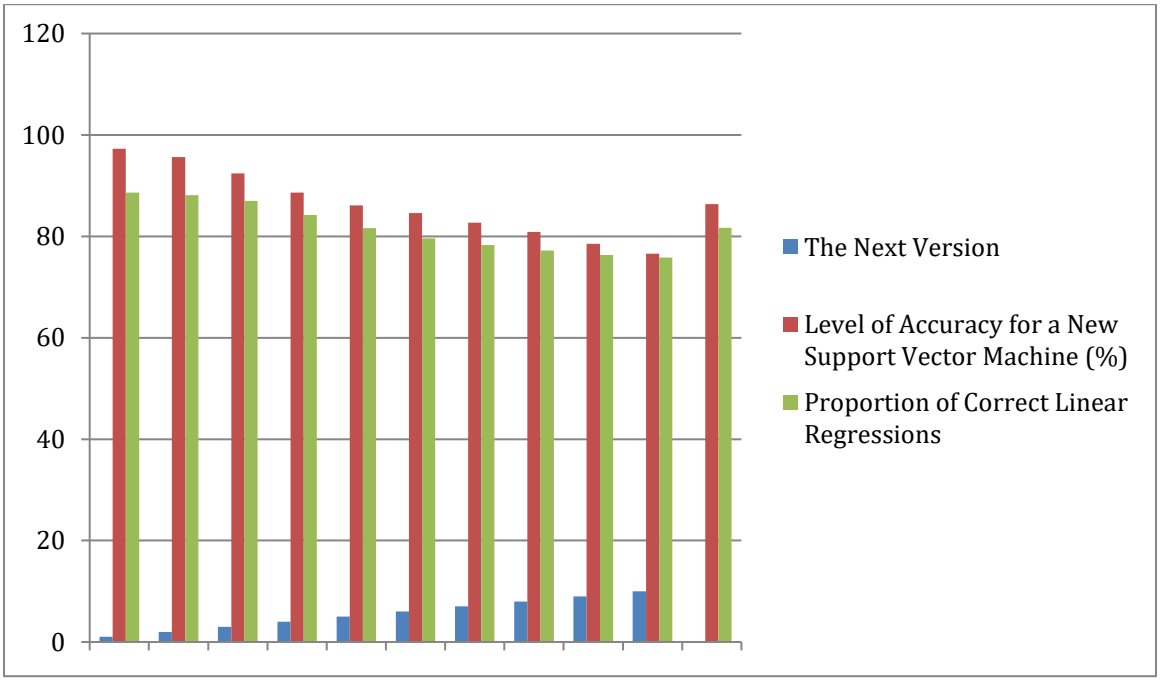

Table 4. Finding independence with a T-test With a p-value of just 0.105 and a dataset confidence interval of 95%, Novel Support Vector Machine outperforms Linear Regression. We utilized a 5% significance threshold since p>0.05.

	Parity of Variances Verified by Levene's Test	**The T-test confirms that**

		F	Sig	t	df	Important (2-tailed)	Average Distinction	Difference in Standard Error	95 Percentage Points on the Distinction	
									Lower	Upper
Accuracy	**Assumed Values for Variances**	1.183	.291	1.707	18	.105	4.6700	2.73591	-1.07793	10.41793
	Not Assumed: Equal Variances			1.707	16.158	.107	4.6700	2.73591	-1.12527	10.46527

Table 5. A group of statisticians conducted an experiment using LR and Novel SVM. By the conclusion of the tenth cycle, both LR and Novel SVM have attained accuracy, dispersion, and average standard deviation. Modern, computerized approaches to linear regression outperformed their human forebears.

Group	N	Mean	Determination of Median	Average with Standard Deviation
Support Vector Machine	10	86.3400	7.07552	2.237498
Linear Regression	10	81.6700	4.97886	1.5445

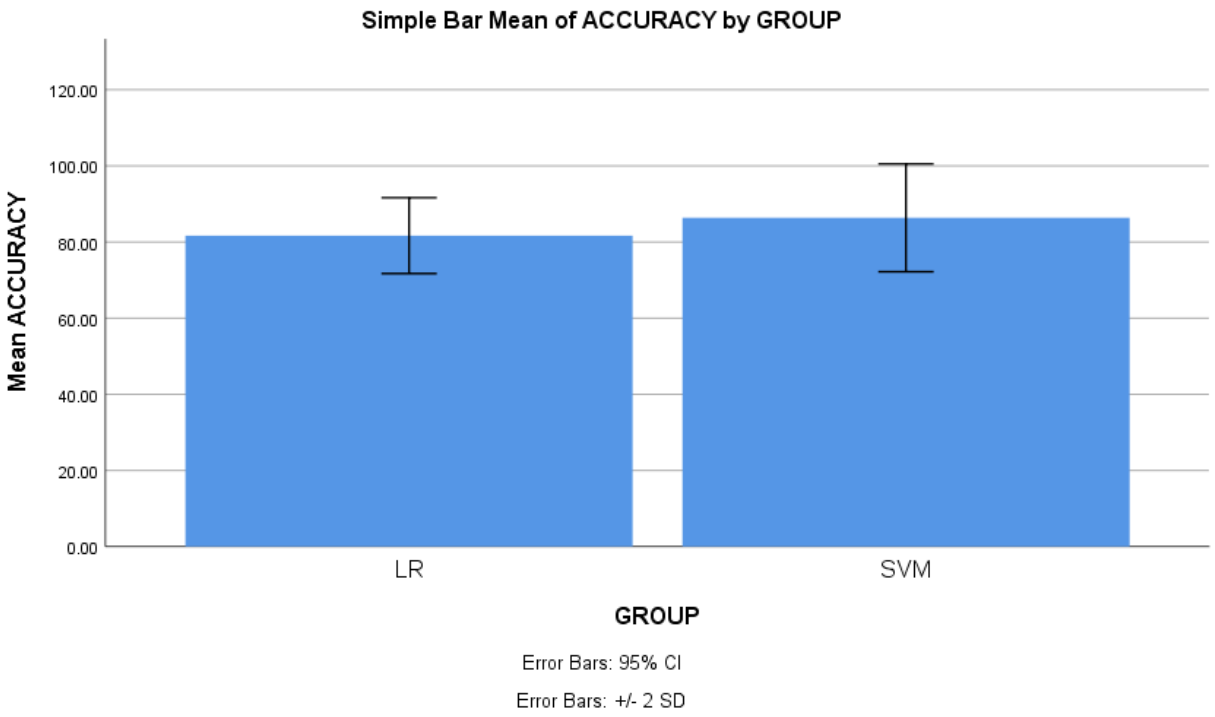

Fig. 1. By contrasting the two approaches side by side, we can get the mean accuracy of LR and Novel SVM. Over LR, Novel SVM Outperforms in Mean and Standard Deviation Comparisons. On the X-axis, the GROUP test displays the results of a comparison between Novel SVM and LR. With a 95% confidence range of +/- 2 standard deviations, the mean detection accuracy is shown on the Y-axis.