E-mail Fraud Detection



Arju Kumar, Saurav Kumar, Kishan Kumar, Bharat Bhushan Naib

Abstract: Spam issues have become worse on social media platforms and apps with the growth of IoT. To address the problem, researchers have proposed several spam detection techniques. Spam rates are still high despite the use of anti-spam technologies and tactics, especially given the ubiquity of rogue emails that lead to dangerous websites. By consuming memory or storage space, spam emails can cause servers to run slowly. One of the most effective methods for identifying and eliminating spam is email filtering. To this end, various deep learning and machine learning technologies have been used, including Naive Bayes, decision trees, SVM, and random forest. Email and Internet of Things spam filters employ various machine learning approaches, and these systems are categorised in this research. Additionally, as more people use mobile devices and SMS services become more affordable, the issue of spam SMS messages is spreading worldwide. This study suggests using a variety of machine learning approaches to detect and eliminate spam as a solution to this problem. According to the trial findings, the TF-IDF with the Random Forest classification algorithm outperformed the other examined algorithms in terms of accuracy. It is only possible to gauge performance on accuracy since the dataset is imbalanced. Therefore, the algorithms must have good precision, recall, and Fmeasure.

Keywords: Convolutional Neural Network (CNN), Onyx Model, Deep Learning, MXNet, TensorFlow, and Face Recognition.

I. INTRODUCTION

In many people's daily lives, mobile phones have supplanted genuine friends. The widespread availability and usage of mobile devices are primarily for sending and receiving SMS messages daily; the industry around this service has ballooned into a multimillion-dollar commercial enterprise. Depending on the country, SMS contributed between 11.3% and 24.7% of the gross national product in 2013.

Manuscript received on 03 June 2023 | Revised Manuscript received on 23 July 2023 | Manuscript Accepted on 15 August 2023 | Manuscript published on 30 August 2023.

*Correspondence Author(s)

OPEN ACCESS

Arju Kumar*, Department of Computer Science and Engineering, SCSE, Galgotias University, Greater Noida (U.P.), India. Email: arju.20scse1010586@galgotiasuniversity.edu.in, ORCID ID: 0009-0004-8899-8975

Saurav Kumar, Department of Computer Science and Engineering, SCSE, Galgotias University, Greater Noida (U.P.), India. Email: saurav.20scse1010931@galgotiasuniversity.edu.in, ORCID ID: 0009-0000-6114-900X

Kishan Kumar, Department of Computer Science and Engineering, SCSE, Galgotias University, Greater Noida (U.P.), India. Email: kishan.20scse1010860@galgotiasuniversity.edu.in, ORCID ID: 0009-0006-2588-0499

Dr. Bharat Bhushan Naib, Department of Computer Science and Engineering, SCSE, Galgotias University, Greater Noida (U.P.), India. Email: bharat.bhushan@galgotiasuniversity.edu.in, ORCID ID: 0000-0002-4238-9697

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an <u>open access</u> article under the CC-BY-NC-ND license <u>http://creativecommons.org/licenses/by-nc-nd/4.0/</u>

One consequence of widespread mobile device usage and the low cost of sending SMS texts is the growth of unwanted bulk communications, especially advertisements. Spam in e-mails is more widespread than in text messages. SMS spam may not be as widespread as e-mail spam, but it still has the potential to aggravate mobile phone users and contribute to social unrest. Spam calls to mobile phones might occur more or less often, depending on your location. Unwanted messages, commonly referred to as "spam," are flooding our messaging system. Text messages sent to mobile phones without prior consent are known as mobile phone spam or SMS spam. Spammers often distribute them in large batches to several targets. Organisations usually resort to this type of spamming, with spreading awareness being one of their primary objectives.

Businesses often send spam text messages to consumers to advertise their goods and services; however, these messages also constitute a security risk due to the ease with which scammers may access users' private information. Spam is the most common and possibly harmful assault on e-mail systems. Unsolicited bulk email or other forms of communication transmitted through an unsecured channel are collectively referred to as "spam." These emails are annoying and potentially dangerous because they contain malware or links to suspicious websites. Breakdowns in safety. Recently, it has become fashionable to utilise machine learning techniques, such as Random Forest and TF-IDF, for spam categorisation. Malicious e-mails with attachments or links may compromise user data and cause server slowdowns. Email spam is on the rise, and organisations must evaluate all the tools they have to combat it.

Individual and corporate victims of e-mail fraud may suffer severe repercussions. For instance, phishing attempts may result in lost money, stolen identities, and ruined reputations. Email fraud can have a devastating impact on businesses, including the theft of sensitive information and a decline in consumer confidence. Fraudsters' constantly shifting methods make it very difficult to identify bogus emails. These bad guys always devise new ways to circumvent established defences and prey on unsuspecting victims. The sheer volume of spam emails makes it difficult, if not impossible, to manually identify and respond to each one. Experts in the field have responded to these issues by developing many methods for identifying fraudulent e-mails [1]. These strategies use cutting-edge tech to detect and stop spam emails, including machine learning, data mining, NLP, and sender reputation monitoring. To prevent individuals and businesses from falling prey to email scams, it is essential to distinguish between authentic and fraudulent messages.

Published By: Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) © Copyright: All rights reserved.



1

II. LITERATURE SURVEY

In a word, yes. These unsolicited text messages and e-mails aim to advertise goods and services or lure recipients into doing fraudulent actions [1]. SMS spam may be restricted in character count, but it poses a financial risk since the receivers may have to pay to process the transaction. As a result, it is crucial to refine spam filtering strategies for both SMS and email to reduce the frequency with which these messages are sent. In SMS spam categorization, a "good word attack strategy" is a tactic employed by spammers to avoid detection by making the spam message seem more like an honest communication by sneaking useful phrases into it [2]. When a classifier encounters an inserted term often used in authentic texts, it may have trouble telling the two apart.

The authors of this study propose a feature reweighting strategy to address this issue by assigning less importance to short-word qualities. This approach aims to reduce the weight given to the added words to reduce their effect on the classifier's output [3]. As a result, the classifier becomes more resistant to well-planned and well-performed word attacks.

To put this reweighting of features into practice, the authors also provide a new rescaling function. This function adjusts their weights according to the inverse of their length to reduce the importance of qualities that are likely to be introduced words [4]. This allows the classifier to focus more on the actual characteristics of spam communications, while the added terms have a lesser effect.

The machine-learning approaches used by Striatal, Mujtaba, and Yasin for SMS spam filtering and detection are somewhat dissimilar. Message size, frequently occurring monograms and diagrams, and Mujtaba and Yasin's use of raw text messages, message length, and information gain matrix as characteristics to detect spam messages, are inferior to their use of message class, indicating that they adopt a more holistic approach. It would be fascinating to compare the efficacy of various algorithms in the context of SMS spam filtering [5]. Keep in mind that the algorithm you use might have a significant impact on how well and how quickly your classifications are made.

There are fewer characteristics. It may be used to evaluate SMS versus email correspondence; overall, the study demonstrates that SMS spam detection is a challenging task. However, machine learning techniques may still help spot SMS spam; therefore, exploring this field is crucial for solving the rising issue of mobile phone spam [6].

III. PROPOSED APPROACH





A. Model / Functionality

The thought fascinates me. It is crucial to continue studying the application of machine learning to the issue of spam filtering in IoT and email networks. It's great that the paper takes stock of the approaches and sensibly arranges them. Machine learning approaches and techniques are employed, utilising several well-known machine learning algorithms due to their ability to process and draw conclusions from large datasets, which include Naive Bayes, decision trees, SVMs, and random forests. Seeing which specific approaches and algorithms are presented and how they compare in terms of efficiency and productivity would be an exciting addition to the research.

The procedure of preparing SMS text messages for use with machine learning methods is also covered in this study [7]. At this point, Keyword symbols may be used to arrange better data that is currently unstructured.

Published By: Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) © Copyright: All rights reserved.



Retrieval Number:100.1/ijese.B77970712223 DOI: <u>10.35940/ijese.B7797.0811923</u> Journal Website: <u>www.ijese.org</u>



The study improved readability by eliminating extraneous words using an English-specific "stop word list remover." Pronouns and propositions, such as "to" and "your," make up the majority of spam text messages, as seen in the visualisation of word frequencies in SMS messages. The most common words in Ham's writings are also stop words, such as pronouns and conjunctions.

Support Vector Machines (SVMs) are another common approach for text categorisation tasks, such as spam detection. This robust approach can handle non-linear decision boundaries and large feature spaces. Random Forest is an ensemble learning technique that combines the conclusions drawn from several decision trees into a single prediction [8]. Popular because of its reliability and ability to handle imperfect information. Yes, in a word.

SVM (Support Vector Machine) is a robust classification and regression tool in supervised learning. SVM seeks the optimal

hyperplane for data classification by maximising the distance between points in different categories. When doing regression using SVM, one must determine which hyperplane best fits the data and minimizes the discrepancy between the predicted and observed values. Text classification is another widespread application of support vector machines (SVMs) following their widespread adoption in bioinformatics and image analysis.

You got it! Specifically, SVM seeks to identify the optimal hyperplane for classifying data with n dimensions (the number of characteristics) into two groups. Hyperplane selection for maximum efficiency for partitioning the population into two subsets [9]. Assuming the hyperplane has been located, further data points may be classified according to whether they fall on the positive or negative side of the plane.



IV. MODEL BUILDING

Fig. 2. Model Building Approach

A. Exploratory data analysis (EDA)

Data scientists use EDA to examine data sets, identify patterns within them, and form conclusions about those patterns and the data as a whole, often by presenting facts visually. Data scientists may use it to determine how to optimise data sources to obtain the necessary information, which aids in identifying patterns, detecting outliers, testing hypotheses, and verifying assumptions.

For researchers to understand what data gathering variables and their interplay reveal, it is crucial to gain insight into implications beyond the scope of the original task. This is where exploratory data analysis (EDA) comes in. You may use it to see how well other methods of statistical analysis hold up, too [10]. The exploratory data analysis (EDA) methods pioneered by American mathematician John Tukey in the 1970s remain popular in the modern era of data discovery.

Retrieval Number:100.1/ijese.B77970712223 DOI: <u>10.35940/ijese.B7797.0811923</u> Journal Website: <u>www.ijese.org</u> Published By: Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) © Copyright: All rights reserved.



E-mail Fraud Detection



Fig. 3. EDA Heat Map Diagram

V. RESULT AND DISCUSSION

Sorry, but as a machine learning language model, I was not provided with the table you referred to. However, I can elaborate on the criteria for evaluating employees' performance.

Accuracy: The proportion of samples for which predictions were correct as a percentage of all models is a measure of accuracy.

Precision is the ratio of real positives (the number of detected samples) to expected positives (the number of pieces that were either correctly or wrongly categorized as positive).

The F-measure is the harmonic mean of recall and precision, representing the proportion of true positives relative to the total number of positives. Samples that should have been considered positive but needed to be [11]. The F-measure is a more accurate indicator of the model's efficacy, as it considers both accuracy and recall.

These metrics are often used to evaluate machine learning models for classification tasks, such as spam detection.

Various methods, such as rule-based systems, content filtering, and machine learning algorithms, are employed to sort through and categorise spam emails [16]. Determining if an email is spam requires a system that uses predefined rules or criteria. With content-based filtering, the words and phrases within an email are examined to determine whether they constitute spam [12]. To categorise new, unlabeled emails and to discover the characteristics and patterns of spam emails, the initial step in using machine learning is to train the algorithms on a large set of labelled emails. The four most common machine learning algorithms may be used for different purposes. Some of the most popular ML approaches for labelling and classifying e-mail spam [17]. These algorithms are educated to recognize spam based on the sender's address, subject line, and message content. The recall, accuracy, and precision of different algorithms are all

examples of performance metrics.

Published By: Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) © Copyright: All rights reserved.





International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319-6378 (Online), Volume-11 Issue-9, August 2023

	Algorithm	Accuracy	Precision	Accuracy_scaling_x	Precision_scaling_x	Accuracy_scaling_y	Precision_scaling_y
0	KN	0.906190	1.000000	0.906190	1.000000	0.906190	1.000000
1	NB	0.974855	1.000000	0.974855	1.000000	0.974855	1.000000
2	RF	0.976789	0.983051	0.976789	0.983051	0.976789	0.983051
3	SVC	0.974855	0.974576	0.974855	0.974576	0.974855	0.974576
4	LR	0.954545	0.959596	0.954545	0.959596	0.954545	0.959596
5	ETC	0.974855	0.959016	0.974855	0.959016	0.974855	0.959016
6	xgb	0.966151	0.955752	0.966151	0.955752	0.966151	0.955752
7	AdaBoost	0.968085	0.948718	0.968085	0.948718	0.968085	0.948718
8	GBDT	0.947776	0.946809	0.947776	0.946809	0.947776	0.946809
9	BgC	0.961315	0.895161	0.961315	0.895161	0.961315	0.895161
10	DT	0.937137	0.854369	0.937137	0.854369	0.937137	0.854369

Fig. 4. Training Data Table

The inbox will export new messages to a dataset in the format below. Whether or not this e-mail is spam depends on how it is analyzed.



Fig. 5. Exported Dataset

Using the methods mentioned above and the likelihood of terms in spam and ham communications, Bayes' theorem and the Naive Bayes Classifier will be employed to determine whether the exported message is spam. The data below reveals the proportion of filtered spam and ham messages [13]. Using the training Using the data, Bayes' theorem, and a Naive Bayes Classifier, we may conclude that the statement "Urgent! Please call 09062703810" is spam based on its similarities to other messages known as spam. In the following example, we use a Naive Bayes Classifier based on Bayes' Theorem to conclude that the message beginning with "Thanx" is, in fact, Spam, since it was exported from the inbox to the dataset.



Retrieval Number: 100.1/ijese. B77970712223 DOI: 10.35940/ijese.B7797.0811923 Journal Website: www.ijese.org



CONCLUSION AND FUTURE WORK VI.

The research was carried out more effectively to take attendance. The deep learning system MobileFaceNet can identify faces in the given dataset with an accuracy of up to 85% when labelled and 90% when recognised. This would benefit management by reducing the time needed to manually take attendance and replacing the RFID card system, which assigns each student a unique identification. Hence, the possibility of card loss won't impact student attendance, and fraudulent attendance will be reduced [14]. Therefore, when the termination is implemented, the institution will profit from decreased enrollment and fewer cases of illegal entry [15]. As future work, the innovative attendance system can be designed to track attendance for all students in a class, and an IoT-connected camera can take snapshots of students and mark their attendance.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.		
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.		
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence.		
Availability of Data and Material/ Data Access Statement	Not relevant.		
Authors Contributions	All authors have equal contributions to this article.		

REFERENCES

- Identification of Spam E-mail Using Information from E-mail Header, 1. Shukor Bin Abd Razak and Ahmad Fahrulrazie Bin Mohamad, 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013.
- 2 The Two Mo's: Mohammed Reza Parsei and Mohammed Salehi. Email Spam Detection Using Part-of-Speech Tagging, 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI).
- Tareek M. Pattewar and Sunil B. Rathod. Using a Bayesian classifier to analyse email for spam content was the topic of a presentation at the 2015 IEEE International Conference on Computer Security and Privacy.
- The authors are Aakash Atul Alurkar, Sourabh Bharat Ranade, Shreeya 4. Vijay Joshi, Siddhesh Sanjay Ranade, Piyush A. Sonewa, Parikshit N. Mahalle, and Arvind V. Deshpande. "A Proposed Data Science Approach for E-mail Spam Classification using Machine Learning Techniques," 2017.
- 5. Together, Kriti Agarwal and Tarun Kumar, in the 2018 proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), described their method in the paper "E-mail Spam Detection using an integrated approach of Naive Bayes and Particle Swarm Optimisation."
- 6. Hezha M. Tareq Abdulhadi, Cihan Varol, At the December 2018 International Congress on Big Data, Deep Learning, and Countering Cyber Terrorism, researchers compared several string-matching algorithms for use in detecting spam emails.
- Xu, Dong, Ivor W. H. Tsang, and Lixin Duan. A method for domain adaptation based on domain-dependent regularisation using numerous independent domains. Neural Networks and Learning Systems, Volume 23, Issue 3 of the IEEE Transactions on (2012).
- Ghulam Mujtaba et al. "E-mail classification research trends: Review and open issues." IEEE Access 2017 is now available. https://doi.org/10.1109/ACCESS.2017.2702187
- Mr. Shrawan Kumar Trivedi. "Spam Detection Classifiers as a Case 9. Study in Machine Learning" International Symposium on

Retrieval Number: 100.1/ijese.B77970712223 DOI: 10.35940/ijese.B7797.0811923 Journal Website: <u>www.ijese.org</u>

Computational and Business Intelligence (ISCBI), 2016 IEEE, 2016. To you, Wanqing, and the others. Naive Bayes Classification for Spam Filtering Via Web Services. Conference on Big Data Computing Services and Applications (BigDataService) 2015, the First IEEE International Conference on. IEEE, 2015.

- 10. S. B. Rathod and T. M. Pattewar. "A Bayesian classifier for contentbased spam detection in electronic mail." 2015 IEEE International Conference on.
- 11. Esra Sahn, Murat Aydos, and Fatih Orhan wrote an article titled "Spam/ham e-mail classification using machine learning methods based on bag of words technique." For 2018, the SIU will host the 26th Annual Conference on Signal Processing and Communications Applications. IEEE, 2018. https://doi.org/10.1109/SIU.2018.8404347
- 12. To cite this article: V. Bhalla, T. Singla, A. Gahlot, and V. Gupta, "Bluetooth-based attendance management system," International Journal of Innovations in Engineering and Technology, vol. 3, no. 1, pp. 227-233, 2013.
- 13. Present and accounted for: Increasing student attendance via parental and community participation, by J. L. Epstein and S. B. Sheldon, The Journal of Educational Research, vol. 95, no. 5, pp. 308-318, 2002. https://doi.org/10.1080/00220670209596604
- 14. "Socioeconomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure," by D. D. Ready, published in Sociology of Education, vol. 83, no. 4, pages 271-286, 2010. https://doi.org/10.1177/0038040710383520
- 15. Ghulam Mujtaba et al. "E-mail classification research trends: Review issues." and open Five (2017)IEEE Access. https://doi.org/10.1109/ACCESS.2017.2702187
- 16. Web Information Retrieval Models, Techniques, and Issues: Survey by J. N. Singh, P. Johri, A. Kumar, and M. Singh. Presented at the 2nd International Conference on Advanced Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022.
- 17. C. P. McCluskey, T. S. Bynum, and J. W. Patchin, "Reducing chronic absenteeism: An assessment of an early truancy initiative," NCCD news, 214-234, vol. 50, 2, 2004. no. pp. https://doi.org/10.1177/0011128703258942

AUTHORS PROFILE



Arju Kumar, I belong to the Supaul district in Bihar. I am a student of B.Tech Computer Science and Engineering at Galgotias University. Currently, I am living in Greater Noida, Uttar Pradesh, Pin: 201310. I have completed my 10th and 12th grades from the BSEB. Currently, I'm pursuing a BTech in the CSE branch at Galgotias University. I possess skills in Java, Python, and Flutter at a basic level, having completed

numerous projects related to Web and Android development. My achievements include participating in programs and contests such as the Smart India Hackathon, NASA Space Research Challenge, Coding Ninja contest, and GFG weekly contest, among others. My hobbies are sketching, watching movies and video editing.



Saurav Kumar, I am from the Nalanda district in Bihar. I am a student of B.Tech Computer Science and Engineering at Galgotias University. Currently, I am living in Greater Noida, Uttar Pradesh, Pin: 201310. I have completed my 10th & 12th from CBSE. Currently, I'm pursuing a BTech degree in the CSE branch at

Galgotias University. I have learned C, C++, Python, and Java at a basic level, but I have also completed numerous projects in Android development. My achievements include participating in programs and contests such as the Smart India Hackathon, NASA Space Research Challenge, Coding Ninja contest, and GFG weekly contest, among others. My hobbies are photography and video editing.



Kishan Kumar, I belong to the Supaul district in Bihar. I am a student of B.Tech Computer Science and Engineering at Galgotias University. Currently, I am living in Greater Noida, Uttar Pradesh, Pin: 201310. I have completed my 10th and 12th grades from the BSEB. Currently, I'm pursuing a BTech in the CSE branch at Galgotias University. I possess skills in Java, Python, and Flutter

at a basic level, having completed numerous projects related to Web and Android development. Mv achievement is that I have participated in programs and contests

Published By: Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) © Copyright: All rights reserved.



6



like:

Smart India Hackathon, NASA space research challenge, Coding Ninja contest, GFG weekly contest and many more like this. My hobbies are sketching, watching movies and video editing.



Dr. Bharat Bhushan Naib, I am from New Delhi, Delhi, India. I am an Associate Professor at Galgotias University, where I teach B.Tech Computer Science and Engineering. Currently, I am living in Greater Noida, Uttar Pradesh, India. I have held various positions in multiple organisations, including the Society for Promotion of Industrial Development and Engineering Research, Samarth

Smart Sustainable Solutions and Services, and K.R. Mangalam University. Currently, I am working as an Associate Professor at Galgotias University. My achievement is that throughout the year, I have honed my quality skills and successfully achieved my goal. I have also led a team of many students as a project guide.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Retrieval Number:100.1/ijese.B77970712223 DOI: <u>10.35940/ijese.B7797.0811923</u> Journal Website: <u>www.ijese.org</u>