How to Cite:
Singh, M. C., Sumanth, P., Sathyanarayana, S. B., & Rithika, G. (2022). Phishing email detection using deep learning algorithms. *International Journal of Health Sciences*, *6*(S3), 8130–8139. https://doi.org/10.53730/ijhs.v6nS3.7944

Phishing email detection using deep learning algorithms

M. Chathar Singh
Asst Professor, Geethanjali College of Engineering and Technology, TS,Hyd, India-501301
*Corresponding author email: chatharsingh.ece@gcet.edu.in

P. Sumanth
UG Students, Geethanjali College of Engineering and Technology, TS, Hyd, India-501301
Email: 19r15a0414@gcet.edu.in

S. Bala Sathyanarayana
UG Students, Geethanjali College of Engineering and Technology, TS, Hyd, India-501301
Email: 18r11a04d4@gcet.edu.in

G. Rithika
UG Students, Geethanjali College of Engineering and Technology, TS, Hyd, India-501301
Email: 18r11a04b0@gcet.edu.in

***Abstract*---**E-mail is one of the most widely recognised channels of communication, whether for personal or corporate purposes. The worst part about spam emails is that they intrude on user's privacy without their consent, and the constant bombardment of spam mails fills up the user's entire email space. Additionally, the issue of wasting network capacity and time checking and deleting spam mails makes it an even more concerning issue. Although confrontation tactics are constantly being upgraded, the results of those methods are now unsatisfactory. Furthermore, phishing emails have been on the rise in recent years. To combat the issue of phishing emails, more effective phishing detection technology is required. We intend to create a phishing email detection tool that analyses the email structure first, using an upgraded Convolutional Neural Networks model with multilayer vectors and Long Short-Term Neural Network (LSTM) to model emails at the email header, character level, email content, and word level all at the same time. To assess the efficacy, we'll use an unbalanced dataset with actual phishing and genuine email ratios. The total accuracy of the experiment achieves a high percentage.

Manuscript submitted: 27 March 2022, Manuscript revised: 9 April 2022, Accepted for publication: 18 May 2022

Meanwhile, this ensures that the filter can reliably identify phishing emails while filtering out as many real emails as possible. Results with help of datasets shows the effectiveness of the proposed approach with respect to the accuracy and F1-score. This encouraging finding outperforms previous detection approaches and demonstrates the efficiency of phishing email detection.

***Keywords*---**phishing, e-mail, data set, spam, convolutional neural networks, long short-term neural network.

Introduction

A digital letter delivered over the internet is known as electronic mail (e-mail). Documents, photographs, music, and videos can all be attached to e-mails and transferred to the recipient's computer. Every day, billions of e-mails are sent around the world due to their low cost and convenience of use. E-mail accounts can be obtained for free or for a fee from a variety of websites. Phishing technique is used to steel personal information using fake email messages and for the purpose of identifying the theft, this is done by sending emails to gain access to personal confidential information. Phishing email crime is increasing very fast for stealing personal information. The person who response the receiving email, or enter the personal information into email then the data of the person is at risk.

In 2018, 236 billion emails were sent everyday with 53.5 percent of them being spam. This year, 14.5 billion spam emails were sent out on a daily basis on average. Spam emails are estimated to have cost businesses $12.5 billion in lost revenue. The annual damage is projected to be over $20.5 billion USD. Spam is defined as "the use of electronic messaging networks to deliver unsolicited bulk messages, particularly mass advertisement, malicious links, and other harmful content." Unsolicited signifies that you received messages from sources without asking for them. So, if you don't recognise the sender, the message could be spam. When people download free services, software, or update their software, they often don't realise they've signed up for those mailers. "Ham," as described by Spam Bayes in 2001, is defined as "emails that are not usually requested but are not considered spam."

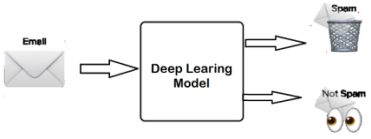


Fig 1. Classification into Spam and non-spam

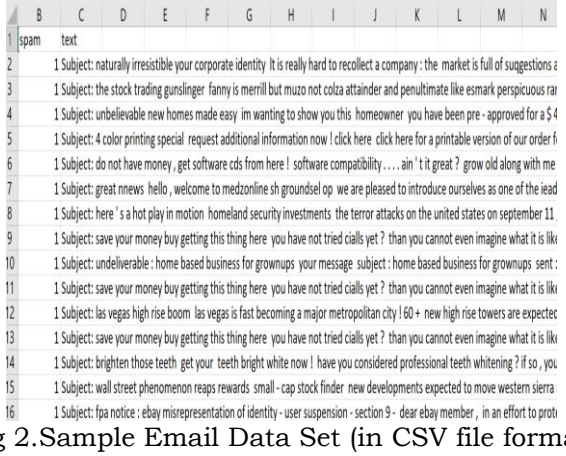
Deep learning algorithms are more efficient since they use a set of training data, which in this case is a batch of pre-classified emails. Many algorithms for email filtering can be found in deep learning approaches. Nave Bayes, support vector machines, Neural Networks, and K-nearest neighbour are a few examples.

Literature Survey

Finding a spam email detection method is a top priority for researchers. We divide the studied literature into two groups in our literature review: deep learning-based spam detection approaches and non-deep learning-based spam detection approaches. MEENA [1] suggested a technique that uses Phishzoo and MLAPT to detect and prevent phishing emails. Phishzoo is a strategy for detecting phishing websites based on their aesthetics, while MLAPT is a technique for stopping phishing emails from being delivered to the system. The user can keep their personal information on the social networking system with the use of these two ways. In a quick presentation, Henry [2] discusses some of the strategies used by phishers to assault users. Furthermore, various strategies have been developed to identify phishing and protect user data from theft. As a result, the focus of this study is on strategies for detecting Phishing attacks. Chhikara [3] provides a quick overview of phishing, its attacks, and the procedures that customers can take to protect their confidential information. This document also includes a phishing overview provided by net craft. On Text and Email Spam datasets, Jain [4] uses SVM, Naive Bayes, ANN, k-nearest neighbours (k-NN), random forest (RF), and LSTM methods. The LSTM algorithm achieves the highest performance. According to the accuracy, recall, precision, and F-measure metrics, this result is 99.01 percent accurate, 99.35 percent recall, 98.74 percent precision, and 99.24 percent F-measure.

Data Set Description

This model utilises email data sets from a variety of internet sources, including Kaggle and sklearn, as well as data sets developed by the author. To train, Kaggle's spam email data set was employed.For getting information, we use our model and another email data set.There are 557 lines and 2 to 3 columns in the data set which is in the form of csv file.574,1001,956 lines of email data are contained in and other data sets isformatted in text format.



	B	C	D	E	F	G	H	I	J	K	L	M	N
1	spam	text											
2		1 Subject: naturally irresistible your corporate identity It is really hard to recollect a company : the market is full of suggestions a											
3		1 Subject: the stock trading gunslinger fanny is merrill but muzo not colza attainder and penultimate like esmark perspicuous ra											
4		1 Subject: unbelievable new homes made easy im wanting to show you this homeowner you have been pre - approved for a $ 4											
5		1 Subject: 4 color printing special request additional information now ! click here click here for a printable version of our order f											
6		1 Subject: do not have money , get software cds from here ! software compatibility . . . . ain ' t it great ? grow old along with me											
7		1 Subject: great nnews hello , welcome to medzonline sh groundsel op we are pleased to introduce ourselves as one of the iead											
8		1 Subject: here ' s a hot play in motion homeland security investments the terror attacks on the united states on september 11											
9		1 Subject: save your money buy getting this thing here you have not tried cialls yet ? than you cannot even imagine what it is lik											
10		1 Subject: undeliverable : home based business for grownups your message subject : home based business for grownups sent :											
11		1 Subject: save your money buy getting this thing here you have not tried cialls yet ? than you cannot even imagine what it is lik											
12		1 Subject: las vegas high rise boom las vegas is fast becoming a major metropolitan city ! 60 + new high rise towers are expectec											
13		1 Subject: save your money buy getting this thing here you have not tried cialls yet ? than you cannot even imagine what it is lik											
14		1 Subject: brighten those teeth get your teeth bright white now ! have you considered professional teeth whitening ? if so , you											
15		1 Subject: wall street phenomenon reaps rewards small - cap stock finder new developments expected to move western sierra											
16		1 Subject: fpa notice : ebay misrepresentation of identity - user suspension - section 9 - dear ebay member , in an effort to prot											

Fig 2.Sample Email Data Set (in CSV file format)

Methodology and Implementation

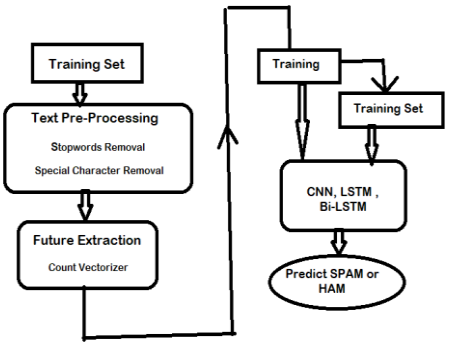


Fig 3. Process of Implementation

The aim of the proposed methodology is to filter the emails into two classes there are Ham and Spam. The following are the steps that make up our methodology: Text Pre-processing, Feature Extraction, Training and Testing the model.

Text Pre-processing

We do data pre-processing before identifying ham and spam communications in order for the suggested model to work more efficiently. We examine each model after reading the email, a set of pre-processing processes that ensures a well-trained model. The following steps are taken to achieve this:

- We take the subject's first line, lowercase it, and remove stop words, special characters, and numerals.
- For the body, we convert each line to lowercase and remove stop words, special characters, and numerals before appending all lines together.
- We then combine the subject and body lines.
- We then go to the feature extraction step, where we delete any words in each email that do not contain any characters or digits.
- The stop words are also deleted to improve the model's accuracy.

Feature Extraction

Because these algorithms can't handle emails in their raw form, which is made up of characters and numbers, feature extraction is a necessary step that seeks to remove any noise that could make our forecast more complicated and inaccurate. The next step is to convert each email into a numerical value vector. This is accomplished by counting the number of times the unique words appear in the given email.

- Count Vectorizer: In this approach each email is represented by a vector of numerical values having each entry represent the frequency of the given word in that specific email

- TF-IDF (Term Frequency-Inverse Document Frequency): The earlier count vectorizer features have the disadvantage of overweighing meaningless repeated phrases like stop words. The Term Frequency Inverse Document Frequency is an alternative that addresses the count vectorizer's shortcomings (TF-IDF). The Term Frequency indicator indicates how often each word appears in each email.

Training and Testing the model

This means, determining the correctness of your model is to train/test it. The data set is divided into two parts: a training set and a testing set. Training takes up 80% of the data, while testing takes up 20%. With the training set, you train the model. Using the testing set, you validate the model.And the block diagram for the approach is shown below.

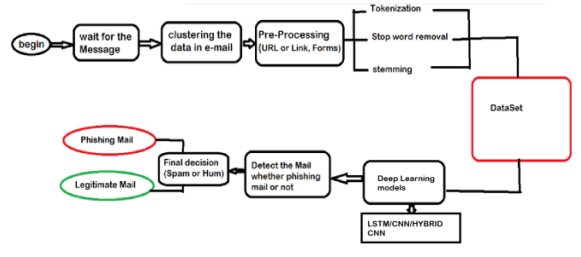


Fig 4. Block diagram of the System

Software Environment

- Pyhton: Python is a high-level programming language for data science and deep learning methods. In this project, we employ Python and its libraries, such as Numpy, Scipy, Pandas, and Matplotlib, as well as TensorFlow and Keras frameworks.
- Django: Django is a high-level python web framework that encourages rapid development and clean, pragmatic design.
- Libraries Used:
  - TensorFlow: It is a free and open-source software library for dataflow and differentiable programming that may be used to solve a variety of problems. It's a symbolic math package that's also utilised in neural networks and other deep learning applications.
  - Numpy: It is an array processing package that can be used for a variety of tasks. It includes a multidimensional array object with outstanding performance as well as facilities for interacting with these arrays.
  - Pandas: It is a free Python library that uses strong data structures to provide high-performance data manipulation and analysis.
  - Matplotlib: It is a Python 2D plotting package that generates high-quality figures in a range of hardcopy and interactive formats among platforms.
  - Keras: It is an open-source high-level Neural Network library written in Python that may be used with Theano, TensorFlow, or CNTK.
  - Scikit-learn: This provides a standard Python interface for a variety of supervised and unsupervised learning techniques.

  - HTML:HTML is a markup language that is used to produce electronic texts (known as pages) that are displayed on the Internet. Hyperlinks are a type of link that allows you to jump from one page to another. Every web page you see was created with a specific HTML version.
- CNN Algorithm: A convolutional neural network (CNN/ConvNet) is a type of deep neural network used to evaluate visual imagery in deep learning.
- LSTM: Long Short-Term Memory is a kind of recurrent neural network.The LSTM features a chain structure with four neural networks and various memory blocks known as cells. The cells store information, whereas the gates manipulate memory. There are three entrances: 1. Forget Gate 2. Input gate 3. Output gate.

Model Evaluation

We'll use the testing data to feed into the model we already have to generate a set of indications. These metrics are used to assess the model's effectiveness. In order to test the entire model, we consider the following factors. For evaluation purposes, standard criteria including as precision, recall, accuracy, and the F1 score are used. Generally, The confusion matrix consists of these measurements.

- True Positive (TP): A test result that appropriately detects the condition when it is present.
- True Negative (TN): A test result in which the condition is not detected when the condition is not present.
- False Positive (FP): A test result that identifies a condition when it does not exist.
- False Negative (FN): A test result that fails to detect a condition when one exists.

Below are some examples of evaluation metrics.

- Accuracy is calculated by dividing the number of correct predictions made by the total number of predictions made.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

- Precision is calculated by dividing the total number of positive class values predicted by the number of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

- The number of correct predictions divided by the number of correct class values in the test data is called recall.

$$Recall = \frac{TP}{TP + FN}$$

- The F Measure conveys the balance between the precision and the recall.

$$FMeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

- Confusion matrix is an N x N matrix is used to evaluate the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values to the deep learning model's predictions.

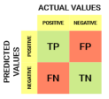


Fig 5. Confusion matrix

Results

The datasets results will be thoroughly analysed. The dataset was used to test various ratios, and the results were compared using different ratios.The estimation of TP and TN should be enormous for the confusion matrix outcome, while the analysis of FP and FN should be minimal.

Fig 6.Home Screen / Interface

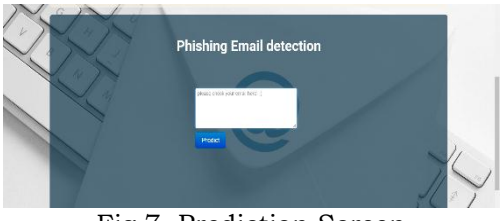


Fig 7. Prediction Screen

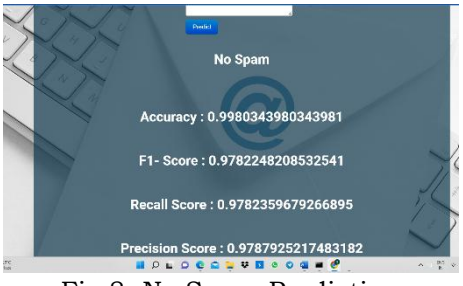


Fig 8. No Spam Prediction

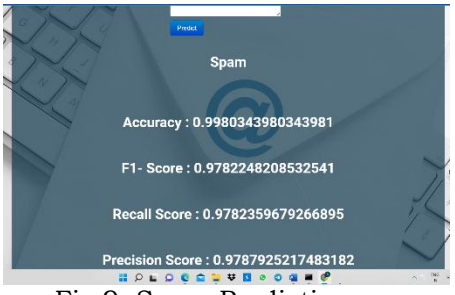


Fig 9. Spam Prediction

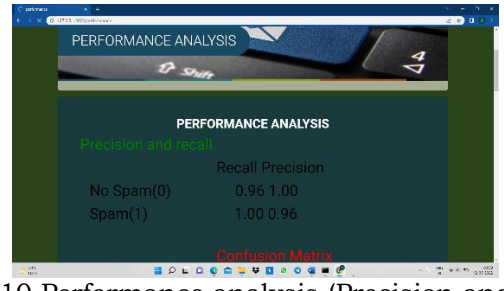


Fig 10.Performance analysis (Precision and Recall)

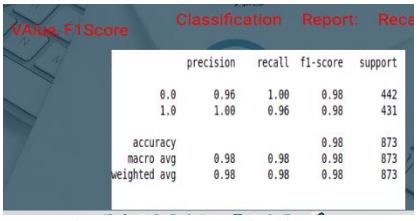


Fig 11. Performance analysis (Comparison)

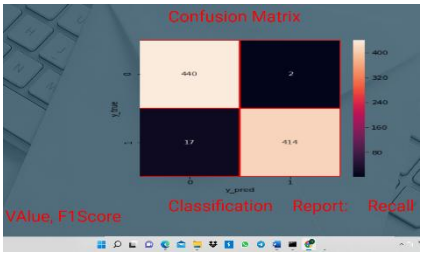


Fig 12. Confusion Matrix

Conclusion

To detect phishing emails, we deploy a powerful deep learning model called. To model the email header and body at the character and word levels, the model uses CNN, LSTM, and Bi-LSTM. As a result, the amount of noise injected into the model is small. The attention mechanism is used between the model's header and body, causing the model to pay greater attention to the more valuable information between them.To conduct tests and evaluate the model, we employ an unbalanced dataset that is closer to the real-world situation. The model produces

a positive outcome. To demonstrate the benefits of the suggested model, several experiments are carried out.

Future Enhancement

For future work, we will focus on how to improve our model for detecting phishing emails with no email header and only an email body.To improve the accuracy of a new phishing detection technique, it is recommended to design a new technology to extract new features from new raw emails. There is a requirement to identify the optimal parameter that improves the accuracy of various classification techniques and provides the best result in spam or ham detection.

References

1. Anti-Phishing Working Group. (2018). Phishing Activity Trends Report 1st Quarter 2018. [Online]. Available:http://docs.apwg.org/Preports/apwg_trends_report_q1_2018.pdf
2. PhishLabs. (2018). 2018 Phish Trends & IntelligenceReport.[Online]. Available: https://info.phishlabs.com/hubfs/2018%20PTI%20Report/PhishLabs%20Trend %20Report_2018-digital.pdf
3. M. Nguyen, T. Nguyen, and T. H. Nguyen, “A Deep Learning Model withHierarchical LSTMs and Supervised Attention for Anti-Phishing,” arXivpreprint arXiv:1805.01554, 2018.
4. A.-P. W. Group et al., “Phishing activity trends report 4th quarter 2016,”USA: Anti-Phishing Working Group (APWG), 2017.
5. A.-P. W. Group et al., “Apwg attack trends report,” USA: Anti-PhishingWorking Group (APWG), 2014.
6. L. M. Form, K. L. Chiew, W. K. Tiong, et al., “Phishing email detectiontechnique by using hybrid features,” in IT in Asia (CITA), 2015 9thInternational Conference on, pp. 1–5, IEEE, 2015.
7. Microsoft, “Microsoft Security Intelligence Report,” https://clouddamcdnprodep.azureedge.net/gdc/gdcVAOQd7/original, 2018.
8. M. Hiransha, N. A. Unnithan, R. Vinayakumar, and K. Soman, “DeepLearning Based Phishing E-mail Detection,” in Proceedings of the 1stAntiPhishing Shared Pilot at 4th ACM International Workshop on Securityand Privacy Analytics (IWSPA 2018) (A. D. R. Verma, ed.), (Tempe,Arizona, USA), 21-03-2018.
9. M. Nguyen, T. Nguyen, and T. H. Nguyen.(2018). ``A deep learning model with hierarchical LSTMs and supervisedattention for anti-phishing.''[Online]. Available: https://arxiv.org/abs/1805.01554
10. Anti-Phishing Working Group. (2016). Phishing Activity TrendsReport 4th Quarter 2016.[Online]. Available: http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf
11. Vishal DineshkumarSoni. (2018). IOT BASED PARKINGLOT. International Engineering Journal ForResearch & Development, 3(1), 9.https://doi.org/10.17605/OSF.IO/9GSAR
12. Anti-Phishing Working Group. (2015). Phishing ActivityTrends Report 1st-3rd Quarter 2015. [Online]. Available:http://docs.apwg.org/ Preports/apwg_trends_report_q1-q3_2015.pdf

13. L. M. Form, K. L. Chiew, S. N. Sze, and W. K. Tiong,Spam e-mail detection technique by using hybrid features, in Proc. 9th Int. Conf. IT Asia (CITA), Aug. 2015, pp. 1_5.
14. M. Hiransha, N. A. Unnithan, R. Vinayakumar, and K.Soman, ``Deep learning-based Spam e-mail detection,'' in Proc.1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secure.Privacy Anal. (IWSPA)A. D. R. Verma, Ed. Tempe, AZ, USA,Mar. 2018