Precise Email Classification using Deep Learning

"Ms. Deepali B. Chavan", "Mr. Suraj S. Redekar"

***Abstract* - Email spam has rapidly expanded in recent years along with the enormous surge in internet users. People are making illegal and immoral use of them. scams conduct, and phishing. It is against the law to send unsolicited emails with dangerous links that could damage our system or try to access your system. The people who are unaware of these scams are the main targets of spammers, who easily construct fake identities and email addresses. In their spam emails, they employ a legitimate name. Therefore, it's crucial to recognize phony spam emails. To accomplish this, this research will make use of machine learning techniques.**

I. INTRODUCTION

The most frequently used method of commercial correspondence is spam email. Despite the availability of other communication tools, email usage continues to rise. Automated email management is critical in today's world, as email volume is increasing regularly. More than 55 perc of all emails are categorized as spam. This exemplifies how spammers squander email users' time and resources while delivering nothing useful. Understanding the various spam email classification strategies and how they work is critical since spammers use complex and imaginative tactics to carry out illicit acts using spam emails. Emails are a popular way of communication for both personal and professional reasons. An intelligent and effective strategy for detecting phishing websites is based on the use of classification or association Data Mining algorithms. To identify phished emails, the suggested solution is a classification issue with two categories: ham and phished. In the area of artificial intelligence known as machine learning, the system is given the capacity to learn on its own without human intervention. For categorization in our model, supervised ML methods are employed.

A. Need of Project:

Email is used in almost every area today, from business to education. Emails are divided into two subcategories: ham and spam. Email spam, sometimes referred to as junk email or unpleasant email, is a sort of email that may be used to hurt any user by stealing their time, computing resources, and important data. Spam email volume is rapidly increasing. Spam identification and filtration are big difficulties for today's service suppliers. Email filtering is one of the most significant and well-known technologies created for identifying and preventing spam. This was accomplished through the use of a variety of ML and DL approaches, such as NB, DT, CNN, and RF. It is demonstrated how machine learning functions in spam detection by providing an in-depth overview of machine learning ideas and approaches.

II. PROBLEM SOLUTION

- Spam content causes missed communication. If such literature contains very sensitive information about any faith, politician, or caste, it may incite severe violence in the area. When spam content begins to propagate, eradicating it from all social media platforms becomes extremely difficult.
- Sensitive or attention-seeking content may include viruses that can gravely harm computers and mobile devices. The directions for future spam filtration and detection research that could be done to increase email platform security and spam detection are addressed.

III. PROPOSED SYSTEM

Dataset - Provide dataset (For a machine that doesn't see data the same way that people do, the acquired data must be standard and intelligible.)

Pre-processing - Real-world data frequently contains sound, absent values, and may even be in an unfavorable format, disqualifying it from being used right away in machine learning models. Preparing the data for a machine learning model by cleaning it is important to increase the model's precision and efficacy.

Feature Extraction: to produce new features from existing ones to reduce the number of features in a dataset (and then discard the original features). This new, more condensed list of features should then act as a summary of almost all of the data contained in the original set of features.

Classification - The Classification algorithm, which uses supervised learning to categorize new observations in light of training data, is used to recognize new observations. In classification, a program makes use of the dataset or observations that are provided to learn how to categorize fresh observations into various classes or groups.

DOI: 10.55041/IJSREM25169

IV. ALGORITHM

A) Decision Tree:

1. A decision tree, a supervised learning method, can be used to address classification and regression issues, but it is often preferred for doing so. It is a tree-structured classifier, where inner nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

2. The Decision Node or Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.

3. The provided dataset's features are used to execute the test or make choices.

4. It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree.

B) Support Vector Machine:
SVM (SVMs, also known as support-vector networks) are supervised learning models with corresponding learning algorithms that examine data used for regression analysis and classification. Given a set of training examples that have all been classified as belonging to either one of two categories or the other, an SVM algorithm builds a model that classifies new examples according to two categories, making it a non-probabilistic binary linear classifier (although there are ways to use SVM in a probabilistic classification setting, such as Platt scaling). An SVM model maps the instances as points in space with as much room as feasible to separate the examples of the different categories. New samples are then projected into that same area and expected to fall into a category based on which side of the gap they fall on.

Two different SVMs Straight SVM:
The term "linearly separable data" refers to data that can be divided into 2 groups using only a single straight line. Linear SVM is used to classify such data, and the classifier utilized is known as the Linear SVM classifier.
Non-linear SVM is used for non-linearly separable data, which indicates that if a dataset cannot be categorized using a straight line, it is non-linear data, and the non-linear SVM classifier is utilized.

C) Convolutional Neural Network:
Convolutional Neural Networks with a focus on image and video recognition applications. CNN is mostly utilized for image analysis applications like segmentation, object detection, and picture recognition.
Convolutional Neural Networks have four different kinds of layers:

1. Convolutional Layer: Each input neuron in a typical neural network is connected to the following hidden layer. Only a small part of the input layer neurons in CNN are connected to the hidden layer of neurons.
2. Pooling Layer: The feature map's dimensionality is decreased using the pooling layer. Inside the CNN's hidden layer, there will be numerous activation and pooling layers.
3. Flatten: Data is flattened when it is made into a one-dimensional array for input into the following layer. We flatten the convolutional layer output to produce a solitary, lengthy feature vector.
4. Fully-connected Layer: The final several layers of the network are known as Fully Connected Layers. The output from the last pooling or convolutional layer is passed into the fully connected layer, where it is flattened before being applied.

V.ADVANTAGES AND DISADVANTAGES

A. Advantages:

- Extraction of link base features from emails.
- Extraction of features based on tags for the entire dataset.
- Word base features extraction.
- The classification outcome of all test results into fishing and normal

B. Disadvantages:

- Needs constant upgrading and maintenance

VI. FUTURE WORK

First, several experiments with and without the mood component are done on a social spam dataset. Following that, we compare the collected results and demonstrate how mood analysis can help to improve the outcomes of social spam filtering. In the future, the model will be evaluated against 1 a large number of standard datasets. This study proposes comparing the presented results to other spam datasets acquired from various sources. Furthermore, on the email spam dataset, further classification and feature algorithms should be tried.

VII. CONCLUSION

Spam detection is crucial for message and e-mail transmission security. Accurate spam detection is a huge task, and several researchers have offered numerous detection approaches. These technologies, however, are incapable of detecting spam correctly and efficiently. We proposed a spam detection technique based on machine learning predictive models to address this issue, the proposed method is more reliable for accurate and timely spam identification, and it will protect message and e-mail communication networks.

REFERENCES

[1] Symantec, 2016 Internet Security Threat Report, 2016 (accessed May 20, 2017).

[2] Kaspersky, Spam and phishing in Q3 2016, 2016 (accessed May 20, 2017).

[3] J. Alqatawna, A. Madain, A. M. Al-Zoubi, and R. AlSayyed, "Online social networks security: Threats, attacks, and future directions," in Social Media Shaping e-Publishing and Academia, pp. 121–132, Springer, 2017.

[4] J. Alqatawna, A. Hadi, M. Al-Zwairi, and M. Khader, "A preliminary analysis of drive-by email attacks in educational institutes," in Cybersecurity and Cyberforensics Conference (CCC), pp. 65–69, IEEE, 2016.

[5] R. A. Halaseh and J. Alqatawna, "Analyzing cybercrimes strategies: The case of a phishing attack," in Cybersecurity and Cyberforensics Conference (CCC), pp. 82–88, IEEE, 2016.

[6] A. Zaid, J. Alqatawna, and A. Huneiti, "A proposed model for malicious spam detection in email systems of educational institutes," in Cybersecurity and Cyberforensics Conference (CCC), pp. 60–64, IEEE, 2016.

[7] A. M. Al-Zoubi, J. Alqatawna, and H. Faris, "Spam profile detection in social networks based on public features," in 2017 8th International Conference on Information and Communication Systems (ICICS), pp. 130–135, April 2017.

[8] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," Expert Systems with Applications, vol. 36, no. 7, pp. 10206 – 10222, 2009.

[9] A. Rodan, H. Faris, and J. Alqatawna, "Optimizing feedforward neural networks using biogeography-based optimization for e-mail spam identification," International Journal of Communications, Network and System Sciences, vol. 9, no. 1, p. 19, 2016.

[10] H. Faris, I. Aljarah, and J. Alqatawna, "Optimizing feedforward neural networks using krill herd algorithm for email spam detection," in Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on, vol., no., pp.1-5, pp. 1–5, IEEE, 2015