International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 10, ISSN No. 2455-2143, Pages 212-216
Published Online February 2022 in IJEAST (http://www.ijeast.com)

SUSPICIOUS E-MAIL DETECTION USING VARIOUS TECHNIQUES

Jeevan Raj S, Raghav K Sejpal, Mrs. Priya N, Dr Mir Aadil
MCA Department
Jain University, Bengaluru, Karnataka, India

***Abstract*— In today's world, email spam has become a serious concern, since the number of internet users has grown rapidly. Illegal and unethical practices, such as phishing and fraud, are taking advantage of the diverse classes of users that use different web services. Users that send unsolicited emails with the intention of disrupting or attracting legitimate customers are known as "spammers" by infecting the user system by sending malicious links in a spam email. Spammers prey on those who are unaware of their deceptions by posing as real people in their unsolicited emails and setting up bogus social media profiles and email accounts. These fraudulent spam emails must be identified. The work is an attempt to analyze different machine learning approaches to serve the purpose. This article uses Deep Learning methods for the identification of spam emails with high precision and accuracy.**

***Keywords*— Spam, Machine Learning algorithms, Deep learning Techniques.**

I. INTRODUCTION

Spam is the practice of sending unsolicited emails or advertisements to a large number of people through the email system. No one has given permission to receive an email if it is unsolicited. " Since the last decade, spam emails have been increasingly common. Overwhelmingly, the internet has been plagued with spam. A waste of space, time, and message delivery speed is spam. Email notification filtering is perhaps the most effective means of detecting spam, however spammers can simply defeat all these misuse detection systems. Until recently, spam sent from specific email addresses could be manually stopped. The arduous and time-consuming chore of manually reviewing and removing junk from emails is tedious. This prompts us to examine the possibility of using data mining to identify email spam. Aside from the attachments, much of the email's content is available in the form of plain old text. Textual mining, a subfield of data mining, is explained in this way. Spam will be detected using a machine learning approach. For junk mail screening, content analysis, whitelist and blacklists of domains (and community-based algorithms) are among the most used options. Spammers frequently utilise text analysis to determine the content of email messages. A wide range of solutions that can be implemented on both the server and customer side are available. They use Naive Bayes, which is one of the most widely used algorithms.

However, in the situation of fake positives, rejecting messages based solely on content evaluation can be a problematic issue. Typically, clients and organizations don't have to worry about legitimate messages being lost. Probably the earliest method for differentiating spam from legitimate content was the use of a boycott. In this method, all messages except those from the specified electronic mail addresses are acknowledged. The "junk mail filtering system" sends an affirmation request to the sender, and if they don't react within a certain amount of time, the message is transferred to a less important queue. This is known as the "white list approach."

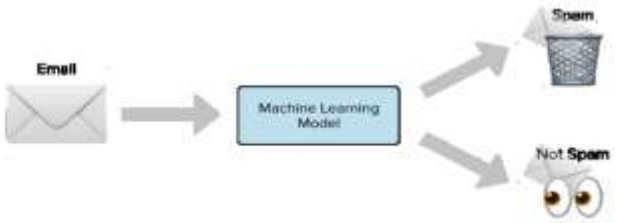

Fig.1. Classification into Spam and non-spam

Rather of relying on trial-and-error methods, machine learning relies on a pre-classified pool of data. Email filtering can take advantage of a wide range of machine learning algorithms. Neural Network models, K-Nearest Neighbor, and Random Forests are just a few examples of these algorithms.

Data mining may be defined as an analytical method that aims to uncover knowledge by studying the patterns in the data. Text mining can make use of a variety of machine learning and natural language processing techniques to glean information from large amounts of text. Email spam detection, online social media network user monitoring for suspicious activity and cybercrime detection are just a few examples of the many uses for text data.

II. LITERATURE REVIEW:

The use of machine learning in email spam identification has been studied in the past. An in-depth literature review of Artificial Intelligence (AI) and Machine Learning (ML) methods for email spam detection was carried out by A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab et al. Two researchers, K. Agarwal and T. Kumar [2].

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 10, ISSN No. 2455-2143, Pages 212-216
Published Online February 2022 in IJEAST (http://www.ijeast.com)

Images and text from the "picture and text dataset for the identification of spam mail with the use of multiple algorithms" have been used by Harisinghaney et al. (2014) and Mohamad &Selamat (2015) et al [4]. KNN, Nave Bayes, and Reverse DBSCAN algorithms were employed by Harisinghaney et al. (2014) [3] to test datasets. OCR library" is used for text recognition, but this OCR does not work adequately. TF-IDF (Term Frequency Inverse Document Frequency) and Coarse pure arithmetic are used by Mohamad & Selamat (2015) [4].

Z. Huang, W. Xu and others (2015) NLP approaches, starting with deep learning's contribution to natural language understanding, have been used to attain [5]. Sentence classification is done using a convolutional neural network (CNN). This study relied on two open-source data sets, both of which had two columns: the email's body text and the label "spam" or "ham." The open source Spambase data collection from D. Dua et alUCI .'s machine learning repository is the first data set to be examined[6]. There are 5569 emails in the data set, and 745 of them are spam. Open source Spamfilter set of data from Kaggle [7] is used in this cond data set, which contains 5728 emails and 1368 spam. Case-based classifiers include k-Nearest Neighbor (kNN) et al. (2007 [8]; 2004) (Trudgian, Dave C. and Sarah Jane Delany, kNN). There is no training step in the kNN method, and instead, training papers are employed for comparison purposes.. It is necessary to find the 'k' number of related papers in order to classify a brand new document. If a class has a large number of documents that are broadly similar, the new document will be placed in that class. A document vector representation of training data is needed to categorise E-mails using kNN. When categorising a new E-mail, the similarity of its document vector to each given document in the training set must be calculated. This is followed by a random selection of classes from among those closest to you in terms of distance (k).

Text mining can be used for sentiment analysis, according to Mouthami et al. (2013)[10]. Negative and positive reviews can be identified using the authors' sentiment fuzzy categorization (SFC) method. Preprocessing, text modification, extraction of features, and classifications are all part of the entire sentiment analysis process. A document-level classification strategy was used to test the algorithm using the Cornell movie review corpus dataset.

Text mining was employed by Olatunji et al (2019) [11] to analyse email spam and the author claimed an efficient accuracy result of 94.06 percent, which was regarded to be greater than other current methods. Various artificial intelligence algorithms have been used to detect bogus news by Ozbay and Alatas (2020) [12]. For the transformation of unstructured information into structured format, text mining is used. 23 different AI approaches have been used to classify the false news datasets, as well. The term frequency (TF) and inverse document frequency frequency (TF-IDF) attribute selection methods were utilised by the writers. The supervised AI-based classifiers have evaluated the selection features for classifying and detecting fake news. Data from BuzzFeed political news, random political news, and ISOT fake news were used in the experiments.

III. METHODS:

Data preprocessing: Data sets having a significant number of rows and columns are always found when the data is examined. Data Preprocessing Steps: Cleaning up the information that has been collected: Here, tasks such as finding and deleting outliers, completing missing values, smoothing noisy data, and "resolving discrepancies" are completed. This step involves the addition of many databases, files, or sets of information. The process of transforming data: Scaling to a certain value necessitates the use of aggregation and normalising. In this phase, the dataset is reduced to a manageable size, but the same analytical results are obtained

3.1 Stop words:

1. There are a number of English words that serve as "stop words," such as "the," "a," "their," "per," "on," "till," and "why." In this case, it is safe to disregard them without forfeiting the meaning of the statement.
2. Tokenization:By using a process known as "tokenization," each email in the corpus is given its own unique token. This is a tally of n emails, each of which has been tokenized and contains m tokens (words). Equation 1 shows how each email is represented.
3. $email_i = (ti_1, ti_2, \ldots, ti_m)$
4. To further decrease the search space, these tokens are examined for the elimination of stop words and stemming procedures.
5. Bag of words “Bag of Words (BOW):

Features can be extracted from documents using this method. Machine learning algorithms can also benefit from these qualities. Each word in the Training dataset is collected into a single Bag of Words vocabulary.

IV. CLASSIC CLASSIFIERS:

4.1. Classification is used to obtain the models that describe the most relevant types of data. An algorithm based on input parameters is used to predict class values. It's possible that applying for a loan could be both harmful and safe. When it comes to data classification, learning and categorization are two independent processes.

4.2. Naïve Bayes:

The Nave Bayes classifier algorithm is a technique of supervised learning that makes use of this technique. On the Bayes theorem, which asserts that features are independent of one another, the Nave Bayes model was based. Spam emails can be classified using the Nave Bayes classifier algorithm, since word likelihood is a major factor in this process. If a

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 10, ISSN No. 2455-2143, Pages 212-216
Published Online February 2022 in IJEAST (http://www.ijeast.com)

word appears frequently in spam but not in ham, it is spam. Naive Bayes classifier method is the most popular method for email filtering today. For the Naive Bayes, the chance of each class is always calculated and then a class with the highest probability is selected as an output. It is always accurate to use Nave Bayes. It is utilized in a wide range of applications, including spam screening, for example.

4.3. Support Vector Machine:

Supervised learning algorithms like the Support Vector Machine (SVM) are popular because they may be used to classify issues in machine learning. It utilizes the notion of decision planes to construct decision boundaries. An object class can be separated by a hyperplane. The hyperplane output of the Support Vector Machine method is used to classify new samples. SVMs using RBF Kernels were employed in this project.

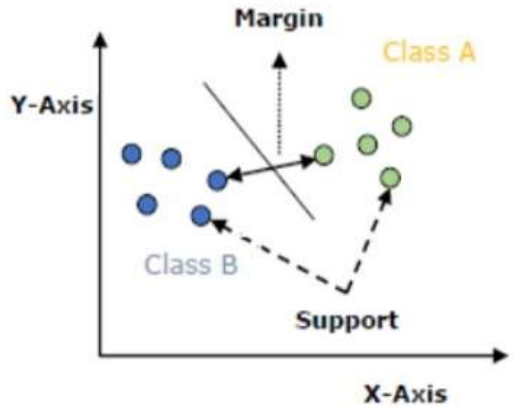

Figure 2: Support Vector Machines

4.4. Decision Tree:

Decision Trees are used to identify spam emails. In spam categorization, a number of different variants of decision tree algorithms were used. For example, the C4.5 tree takes into account the training dataset's attribute values when creating a tree.

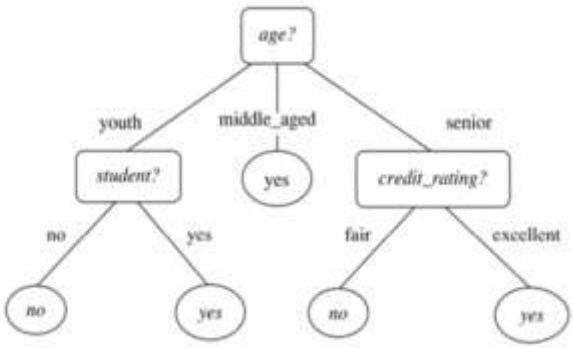

Figure 3: Decision Tree Structure

"Decision tree classifiers" can be built without "any domain expertise or parameter setup that is adequate for examining knowledge." Information in multiple dimensions can be handled by it. Decision tree induction's learning and categorization phases are simple and fast. In order to classify a tuple, a characteristic choice event is used to select the most important feature. An unusually large number of the tree's branches may show signs of turbulence and inconsistencies in the informational collection during the manufacturing process. Tree trimming aims to identify and remove these branches in order to improve the accuracy of a classifier on an otherwise insignificant piece of information.

One attribute's frequency table is used to calculate entropy.

$$E(S) = \sum_{i=1}^{c} -p_i \ log_2 p_i$$

Using the frequency table for two qualities, we can calculate the entropy.

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

4.5 . K- Nearest Neigbour:

There is a supervised classification approach known as K-nearest neighbors. This technique uses a set of data points and a set of data vectors to predict the categorization of a new sample point." The LAZY algorithm is K-Nearest neighbor. In the case of a LAZY algorithm, it will only attempt to memorize the procedure that it was unable to learn on its own. It doesn't come to its own conclusions. Euclidian distance can be used as a similarity measure to classify new points using K-Nearest Neighbor (KNN). Using the Euclidean distance, one can find out who is in the vicinity of a certain point.\

$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$

4.6. Ensemble Learning Methods:

All of the local models and the global model will be accepted as input into the ensemble for making predictions in our proposed work. As a final step, an ensemble theory is implemented to the expected results based on the inputs from the several local models. The ensemble was put together using a process known as bagging.

4.7. Random Forest Classifier:

Decision trees of various shapes and sizes are used in a random forest classifier, which is an ensemble classifier. Data taken at random from the training set in order to construct a decision tree. When splitting at a node in a tree, a random subset of the input features is generated. Ensemble's generalization error (features should not seem the same) can be decreased if randomness is present. Randomization will reduce the co-relation of the decision tree.

4.8. Bagging:

1. A technique known as Bagging has been shown to improve accuracy when used with random subsamples. It

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 10, ISSN No. 2455-2143, Pages 212-216
Published Online February 2022 in IJEAST (http://www.ijeast.com)

works by slicing up the data and then running the function in parallel. This decreases the chance of error and saves time.

2. Bagging is a type of Bootstrap. AGGregatING
3. Because it incorporates the best aspects of both classifiers, a boosting technique has greater advantages. The unclassified dataset samples are likewise given higher weight in the boosting strategy. MNB and J48 Decision Tree classifiers use machine learning methods to do the bagging.

4.9. Boosting And Adaboost Classifier:

Boosting is completed by creating a model from training data, then creating a second model that corrects the original model's errors.

An adaptive boosting technique is called AdaBoot.

AdaBoost was the first successful modelling approach that was used for binary classifications. AdaBoost explains the enhancing.

V. IMPLEMENTATION:

The model is implemented in Visual Studio Code and a dataset from the "Kaggle" website is utilized as a training dataset in this module. For better machine performance, the inserted dataset is first verified for repetitions and null values. As a result, the dataset is divided into two sub-datasets, one of which is referred to as "train dataset" and the other as "test dataset." In the following step, the "train" and "test" datasets are supplied as text-processing arguments. punctuation symbols and words on the stop word list are deleted and returned as clean words in text-processing software "Feature Transform" is then applied to these sanitized words. To construct a vocabulary for the machine, the clean words produced from the text-processing are used for "fit" and "transform." For "hyperparameter tuning," the dataset is used to discover the classifier's optimal value based on the dataset.

The system is fitted with a random state using the values obtained from the "hyperparameter tweaking." In order to test new data, the model's state and its features are preserved. Python classifiers from the sklearn module are used to train the machines with the data from the previous section.

VI. RESULTS:

In order to improve its accuracy, we've trained our model with various classifiers. The user will be presented with the results of each classifier's evaluation. To determine if a piece of data is "spam" or "ham," a user can compare the findings of all of the classifiers. Graphs and tables will be used to explain the results of each classifier. For the purpose of training, the data was downloaded from the "Kaggle" website. The dataset is called "spam.csv." "emails.csv" is the name of a CSV file that contains data that was not utilised in the learning of the machine; it serves as a test. The document is now ready for the templates once the final edits have been made to the text. You can use the Save As command to create a new copy of the template file, and then follow the naming scheme specified by your conference to identify your paper. Select all of the information of the newly formed file and then import the text file you just made. Use the scroll down window on the left of the MS Word Format toolbar to design your paper. There are several Classifiers and their Accuracy Scores listed in Table 1.

Table 1: Accuracy Scores of Various Classifiers

S.No	Classifiers	Accuracy Scores
1	K-Nearest Neighbour	0.90
2	Naïve Bayes	0.86
3	Decision Tree	0.92
4	Random Forest	0.88
5	AdaBoost Classifier	0.93
6	Bagging Classifier	0.92
7	Support Vector Classifier	0.89
8	Support Vector Classifier-RBF	0.94

Figure 4: Accuracy scores

VII. CONCLUSION:

These methods and technologies are examined in this research in order to create Suspicious E-Mail Detection. Some systems are discussed in great depth in regards to the methods they use and the results they produce, as well as the advantages and disadvantages of each. Also, a brief overview of a few studies has been provided.

VIII. REFERENCES:

[1]. Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295. [08907831]. https://doi.org/10.1109/ACCESS.2019.2954791.

[2]. K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

[3]. Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on, pp.153-155. IEEE, 2014.

[4]. Mohamad, Masurah, and Ali Selamat. "An evaluation on the efficiency of hybrid feature selection in spam email classification." In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on, pp. 227-231. IEEE, 2015.

[5]. Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[6]. D. Dua and C. Graff, UCI machine learning repository, 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[7]. karthickveerakumar, Spam filter, 2017. [Online]. Available: https://www.kaggle.com/karthickveerakumar/spam-filter.

[8]. Cunningham, Padraig, and Sarah Jane Delany. k-Nearest neighbour classifiers. Multiple Classifier Systems pp. 1-17. 2007.

[9]. Trudgian, Dave C. Spam classification using nearest neighbour techniques. In Intelligent Data Engineering and Automated Learning–IDEAL pp. 578-585. Springer Berlin Heidelberg, 2004.

[10]. Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013. Sentiment analysis and classification based on textual reviews. In 2013 international conference on Information communication and embedded systems (ICICES), IEEE, pp.271276.

[11]. Olatunji, S.O., 2019. Improved email spam detection model based on support vector machines. Neural Computing and Applications, 31(3), pp.691-699.

[12]. Ozbay, F.A. and Alatas, B., 2020. Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, p.123174.

[13]. Roy, P.K., Singh, J.P. and Banerjee, S., 2020. Deep learning to filter SMS Spam. Future Generation Computer Systems, 102, pp.524-533.

[14]. Naem, A.A., Ghali, N.I. and Saleh, A.A., 2018. Antlion optimization and boosting classifier for spam email detection. Future Computing and Informatics Journal, 3(2), pp.436-442.

[15]. Khamis, S.A., Foozy, C.F.M., Ab Aziz, M.F. and Rahim, N., 2020. Header Based Email Spam Detection Framework Using Support Vector Machine (SVM) Technique. In International Conference on Soft Computing and Data Mining, Springer, Cham, pp.57-65.