Classification of Spam Emails using Deep learning

2021 1st Babylon International Conference on Information Technology and Science (BICITS) | 978-1-6654-1573-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/BICITS51482.2021.9509909

Nuha H. Marza
Department of Computer science
Science College For Women
University of Babylon
Babylon, Iraq
nuha.marza@student.uobabylon.edu.iq

Mehdi E. Manaa
Department of Information Networks
College of Information Technology
University of Babylon
Babylon, Iraq
it.mehdi.ebady@itnet.uobabylon.edu.iq

Hussein A. Lafta
Department of Computer science
Science College For Women
University of Babylon
Babylon, Iraq
wsci.husein.attia@uobabylon.edu.iq

***Abstract*—The Internet has become an integral part of modern life. One of the most critical aspects of the Internet is collaboration. Email is a communication tool that can be used for both personal and professional purposes. Spam messages are not intended to be received by addressee of emails, and therefore are often regarded as unwanted bulk emails. Every day, a wide range of people use email to connect globally. Currently, large numbers of Spam emails are logic genes. Being in large quantities already causes real frustration for both internet users and providers. For instance, it degrades user analysis data, encourages network virus migration, expands stack on arrangement movement, absorbs mail server storage, wastes time and network bandwidth, and depletes the vitality of real emails among the Spam. It is therefore necessary to prevent the spread of Spam. Given the fact that there are several data mining techniques beneficial in preserving security, they can also be of use in classifying Spam email. As for the present work, the Min-hash technique is combined with the Deep Neural Network (DNN) algorithm to classify emails into Spam and Ham. The results indicate that a remarkably high accuracy rate (98%) is obtained by using this combination, which means that it is an effective method to be adopted and further developed in the field of Spam detection and classification.**

Keywords— Spam emails, Data Mining, Classification, Deep Learning, Data Security, Min-hash.

I. INTRODUCTION

Network security is the process whereby a safe environment is provided for computer, users, and programs in order to perform their essential functions. It can be achieved by means of taking both physical and software protective measures to deter unauthorized access, misuse, malfunction, modification, deterioration, or improper disclosure of the underlying network infrastructure. At the heart of information security, there are three key goals to be achieved, namely ensuring confidentiality, durability, and usability [1]. Fig. 1 illustrates the information security triad of CIA (Confidentiality, Integrity, and Availability).

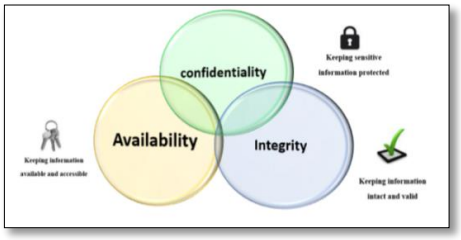

Fig.1. The Confidentiality, Integrity and Availability triad for Information Security

Email has now become one of the easiest and cheapest means of contact. Email popularity, however, has further raised Spam emails over the past few last years. In order to categorize the email as Spam or non-Spam (Ham), data mining classification algorithms are used. Emails is an effective form of online communication because it saves resources and tends to minimize communication time, making it a popular means of communication for private communication and technical communication. Simple data transmission, as well as initial and other files that can be sent worldwide, are supported through company emails. There are also other occasions where numerous attacks affect the emails that users send, which can be active or interactive. Emails are often received from unknown sources, some of which contain meaningless information that is not important or relevant to the recipient. Spam mails are a well-known way of transmitting unnecessary or broad data to a list of particular or random emails addresses. Spam mail can thus be defined as a subset of internet Spam, whereby messages are delivered to all recipients via email, who are linked in some way or other to the same or similar post [2].

Spam emails may also include malware in form of scripts or other files that are executable and can harm the user's system. Most emails and Spam lists are created by thoroughly searching the UseNet and stealing the internet email list. Such emails meet the three key criteria below:

1) *Anonymity:* the sender's address and name are secret.
2) *Bulk mailing:* postal messages delivered to an enormous group of persons.
3) *Unsolicited:* receivers do not request the email.

The aim of this paper is to use the Min-hash technique combined with Deep Neural Networks (DNN) to identify spy communications and eventually prevent the issues that Spam

emails create. The proposed method is built and implemented through the use of data analysis.

II. RELATED WORKS

Various works by authors have previously been published to demonstrate the architectures, techniques, and algorithms under which deep learning is used to classify email, according to a literary study. Nevertheless, the Min- hash technology was not commonly used for email classification, as one dimensional Convolutional Neural Network (1D CNN) only was used to classify emails into Ham and spam.

DNNs have been commonly used in many fields since the development of Deep Learning (DL), which eventually highlighted the critical security issues related to DNNs. Many studies have examined the security of DNNs and they identified several vulnerabilities through proposing a number of attack methods [3], including black-box attacks [4] and white-box attacks [5].

The study presented in [6] makes use of a Deep Neural Network classifier, which is one of the DL architectures, to identify a dataset. They coupled the classifier with the Discrete Wavelet Transform (DWT), a strong feature extraction method, and principal components analysis (PCA). Their results turned out to be very successful throughout all performance steps.

As for [7] , the authors classified emails into Spam and not-Spam (Ham) by analyzing the whole content (i.e. both image and text), and processing it through independent classifiers using Convolutional Neural Networks. They proposed two hybrid multi-modal architectures by forging the image and text classifiers. It is a study approaching our research.

A comprehensive survey is presented in [8], wherein the authors shed light on a broad range of text classification algorithms, such as the Support Vector Machine, Decision Tree, and Rule-based Classifiers.

The authors in [9] present the preliminary results of their study, whereby deep learning is used in legal document analysis. Their experiments were performed on four datasets of actual legal matters, after which their deep learning findings are compared to the results obtained using an SVM algorithm. Their outcomes revealed that Convolutional Neural Network (CNN) performs sufficiently when used with a wider training dataset, and it is therefore found to be a suitable tool for text classification in the legal industry.

As for the studies and experiments presented in [10] [11] [12] [13] , the authors of each work proposed novel strategies for email Spam identification, whereby their work is based on SVM and feature extraction. The application of their proposed strategies on test data sets all resulted in high accuracy rates of about 98%.

III. OVERVIEW OF THE DEEP LEARNING

A general introduction to Deep Learning (DL) involves defining it as sub-field of machine learning that focuses on learning several levels of representations. It is performed through building a hierarchy of features in which lower levels classify higher levels, and lower level features can be used to describe several higher level features [6] [14] , as illustrated in Fig. 2.

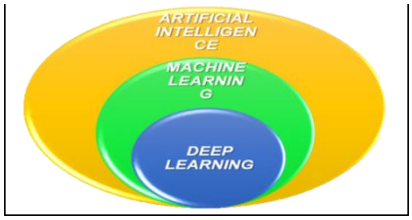

Fig.2. Machine learning and Deep Learning are subfields of Artificial Intelligence.

Different DL architectures exist. A popular type of such architectures is the Convolutional Neural Network (CNN), which has often used as an alternative recently because of their ability of performing complicated operations through the use of convolution filters [6] [15]. A standard CNN structure consists of several fully-connected layers that convert the previous layers' 2D feature maps into 1D vectors for classification, followed by a sequence of feedforward layers that introduce convolutional filters and pooling layers [15] [16].

DNNS are also a type of DL architectures which have been successfully used in classifying and regressing data in a variety of fields. In this type of networks, the information flows from the input layer to the output layer through several hidden layers in a typical feed-forward network (more than two) [17].

Figure 3 shows a standard DNN architecture, which includes an input layer with neurons for input characteristics, an output layer with neurons for output classes, and hidden layers. As for Fig. 4, it illustrates the neural node in detail.

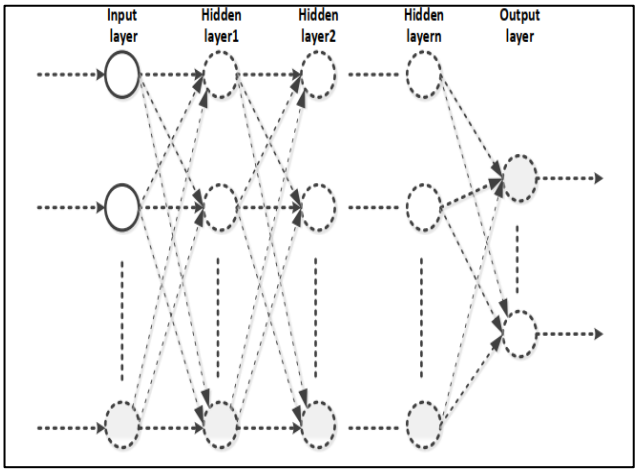

Fig. 3. Structure diagram of deep neural network

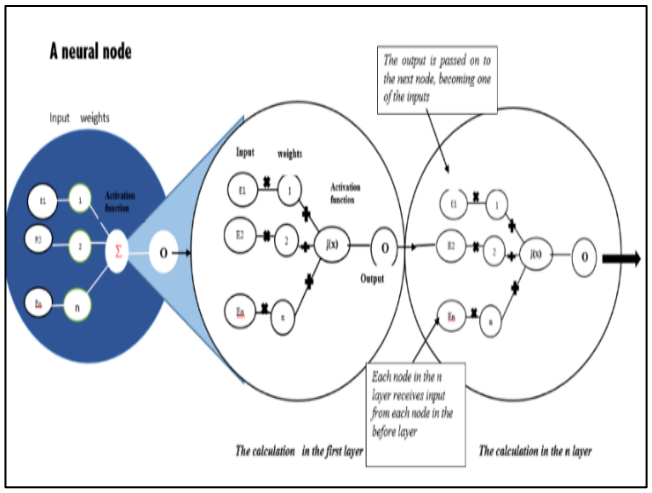

Fig. 4. neural node in Deep Learning.

The DL paradigm extends standard NN by incorporating multiple hidden layers into the network between input and output layers, for modeling more complex and non-linear relations. Due to its outstanding progress in being the perfect approach to a number of problems, this theory has piqued the interest of researchers in recent years [6]. A 1D CNN throughout this work and a five-layer system of hidden layers was proposed. Starting with 12 nodes in the first secret layer and 24 nodes in the second, there were 48 nodes in the third layer, followed by 24 and 12 nodes in the fourth and fifth layers, respectively. 32 batches were set up after the training set was trained.

IV. DATA MINING

The data mining processes and occurrences are statistically important. There are various protocols adopted in data mining and assigned to different data mining techniques. As for the present work, the Min-hash technique is adopted to gather and obtain relevant information. As for the classification, it is a statistical model that displays and generalizes the data into established fixed groups in a predefined manner [18].

A. Min-hash Technique

Min-hash is mostly used in applications that need massive amounts of data. One of the Google techniques [19] is to use it in text similarities. The major stages involved are the k-shingles, min-hash function, and signature matrix.

B. Hashing shingles (k-shingles)

In document similarity, the word K-shingle is often used. Documents are split into a number of tokens depending on the length of k in this technique [20]. Supposing that a text contains the string "This email is Spam and not legitimate" If the value k=3 is used, then the total number of generated tokens is (n-k+1), where n is the total number of words in the documents and k is the shingle length . In the present study, the text files are divided into shingles depending on the number length of k. Table I explains in detail how the group keys are produced, When the k value of k-shingle equals (3).

TABLE I. CHARACTERISTIC MATRIC-BASED K-SHINGLE (K=3)

#Shingling	**Emails**	
	E1	***E2***
This email is	1	1
email is Spam	1	1
is Spam and	1	0
Spam and is	0	1
And is not	1	1
:	:	:
Shingle m	M	m

The algorithm that implements the major steps of k-shingle hashes for emails whereas the input is Emails (E1,E2,….En) and the output is Characteristic Matrix (M); is clarified as follows steps:

1) Preprocess the emails text by.
2) removing the punctuation.
3) removing adjusting the white space.
4) Choose k number
5) Set emails to group based on k
6) Hashing set (shingling)
7) Find existing tokens in emails
8) Generate characteristic Matrix

C. Min-hash Functions (for each shingle)

The Min-hash method takes tokenized text and transforms it to a collection of hash integers, after which it finds the lowest value (minimum). "(1)" gives the general shape of the Min-hash function [21] .

$$h(x)=ax+b \bmod P \quad (1)$$

Where a & b are two random values, X is the hash function value for the tokens and P is the prime number (greater than the maximum number of x) [20].

In this research, the Min-hash method is used to generate the matrix, and the Min-hash steps applied for generating the codes through the use of a single hash function. Table II explains the result of implementing hash function-based k-shingle to characteristic matrix.

TABLE III. VALUES OF CHARACTERISTIC MATRIX WITH MIN-HASH

#Shingling	**Characteristics Matrix**					
	E1	***E2***	***H1***	***H2***	***H3***	***H4***
998816769	1	1	5	2	3	9
351110407	1	1	8	10	10	0

316870923	1	0	1	6	2	6
1976815438	0	1	9	2	7	9
339473466	1	1	6	9	1	1
:	:	:	:	:	:	:
M	M	M	M	m	m	m

The algorithm that implements the main steps Min-hash functions hashing on emails whereas the input is Characteristic Matrix M, and Hash Functions (h1, h2, h3, …, hn) and the output is Signature Matrix (S); is clarified as follows steps:

1) Picking n randomly hashing functions h1, h2, h3, …, hn.

2) Construct the signature Matrix S from characteristic Matrix M, where each row (i) is a hash function and each column (c) is a emails.

3) Then, set SIG(i,c) as signature matrix element for the hash h function and column c.

4) Convert the long bit vector into short signatures.

5) To every column c in documents, do next steps:

a) if c has 0 in both documents' rows r, do nothing.

b) if row has 1, then, for each i=1,2, ……, n set SIG(i,c) to the smaller value of the current value of SIG(i,c) and hi(r) Then Pr[hπ(c1) =hπ(c2)] =sim (c1, c2).

D. Min-hash Signatures

After finding the similarities in the emails, as well as the min hash for all emails, the Min-hash Signatures can be obtained through the construction of the signature matrix by considering each row in the order in which it appears. SIG (i, c) is the signature matrix unit for the ith hash function and column c. Supposing that SIG (i, c) is for both I and c at first ∞, Row r is operated with in the following manner:

1) Calculate h1 (shingles) to hn (shingles).
2) Do the following for each column c:
a) If c has 0 in row, do nothing
b) If c has 1 in row (shingles), then set SIG(i, c) to the smaller of the current value of SIG (i, c), and hi for each I = 1, 2,..., n. (shingles) [21], as shown in Table III.

TABLE III. SIGNATURE MATRIX FOR THE ALL EMAILS

Hash	**Emails**	
	Email 1	***Email 2***
#1	1	5
#2	2	2
#3	1	1
#4	0	0

V. THE PROPOSED METHODOLOGY

The proposed Methodology consist of the fallowing steps:

1) *Data set :* The dataset used in this study can be found on Kaggle, which is a machine learning database. There are 5725 instances in the dataset of "Spam filter", with two columns for class and email string. Fig 5. shows sample of the dataset.

2) *Data cleaning :*Data cleaning is an essential aspect of data science. Working with skewed data will cause a large number of problems. This process entails breaking down into words and dealing with punctuation and case. In this work, unnecessary values are excluded, as is the elimination of stop words, symbols, and punctuation marks, and convert data types are used.

3) *Calculating Hashing shingles (k-shingles):* K-shingles with (k=3) length was used in this work. The characteristics matrix and signature are implemented to generate a dense matrix from large sparse matrix. Characteristic matrix was dense using (h=4) Min-hash function. The values of the min-hash were used to build a signature matrix. These values feed to deep neural network as input vectors.

4) *Calculating Hash crc32 (Shingles)*: After the k-shingle stage, the Hash functions (crc32) is used. As for the k-hashed tokens, the Min-hash algorithm is used.

5) *The data is split into training and testing***.**

6) *The number of hidden layers, nodes on each layer,* training batches, as well as the number of training data for each batch in a DNN with several hidden layers are set.

7) *The DNN Spam classifier is used to classify the checked emails,* so as to decide whether or not they are considered to be Spam. The DNN classifiers can be created by training the DNN. And the output for this stage is Optimum Weights.

8) *The obtained email classification result* are compared with the real tags to ensure that the algorithm is indeed accurate. The result is given as Spam or Ham.

The essential stages of the proposed system are illustrated in Fig. 6.

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1
...	....	....
5721	Subject: re: research and development charges to...	0
5722	Subject: re: receipts from visit jim , thanks ...	0
5723	Subject: re: enron case study update wow ! all ...	0
5724	Subject: re: interest david , please , call sh...	0
5725	Subject: news : aurora 5 . 2 update aurora ve...	0

Fig. 5. Sample of Data set used in this work

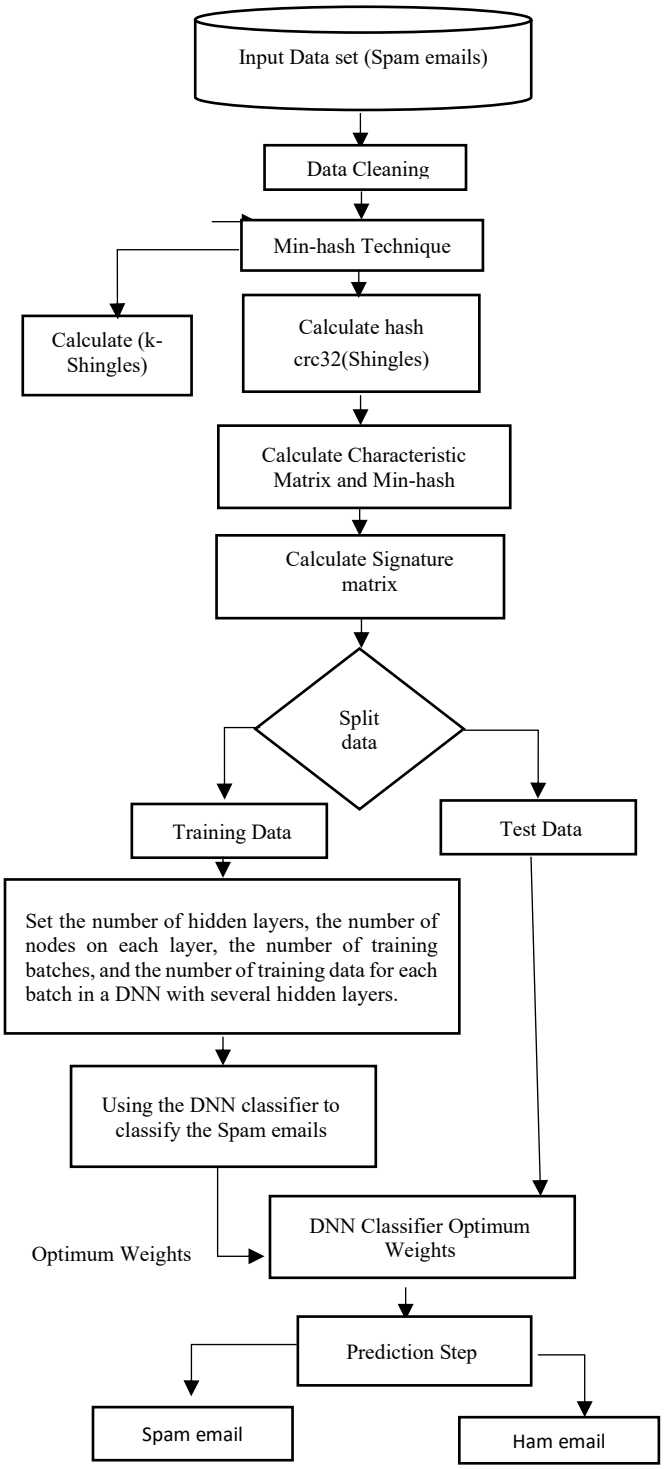

Fig. 6. The main steps of the proposed system.

VI. RESULT AND DISCUSSION

The results of the experiment conducted in this study revealed that DNN with Min-hash function outperformed the use of DNN individually, particularly in terms of email classification accuracy. It was therefore observed that the DNN had a positive classification effect.

Throughout this work, a five-layer system of hidden layers was proposed. Starting with 12 nodes in the first secret layer and 24 nodes in the second, there were 48 nodes in the third layer, followed by 24 and 12 nodes in the fourth and fifth layers, respectively. 32 batches were set up after the training set was trained. The Python environment was used to carry out this work, and 70% to 30% of the data was taken for the purpose of training and testing the results. Whereas previous authors did not use Min-hash with DNN but only used DNN ,and the accuracy is described as follows [22]:

$$Accuracy = \frac{number\ of\ correctly\ identified\ \ email}{total\ nmber\ of\ emails} \quad (2)$$

Or

$$Accuracy\ = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

where, TP, TN, FP, and FN represent the True Positive, True Negative, False Positive and False Negative, respectively [23] The Recall is described as follows:

$$Recall = \frac{(Tp)}{(TP+FN)} \quad (4)$$

Precision is described as follows:

$$\text{Precision} = \frac{(TP)}{(TP+FP)} \quad (5)$$

Table IV is clearly shown that the min-hash with deep neural conducted high results in term of accuracy compared with the authors in [6] and [7] which are satisfied accuracy less 97%.

The main contribution of this work is using the hash to form the sparse characteristics matrix and then using min-hash for generating signature matrix with dense dimensions. The values of this signature matrix is used as an input for feeding deep neural network to get high results in term of accuracy for Ham and spam emails.

TABLE IV. COMPUTATIONAL RESULTS COMPARISON ACCURACY, RECALL AND PRECISION IN DNN& MIN-HASH AND DNN

Method	**Accuracy**	**Recall**	**Precision**
Min-hash +DNN	In proposed system = 98%	The proposed system = 95%	The proposed system = 88%

VII. CONCLUSION

The Min-hash technology was not used with email classification before, as the email was classified using 1D CNN to classify emails into Ham and spam. in this paper presented an email classification algorithm, whereby data mining is used to design and execute an effective method to differentiate and classify email into Spam and Ham. Neural networks have a lot to offer the computer community. Their ability to learn allows them to be very adaptable and strong. Furthermore, there is no need to comprehend the task's internal mechanics. Because of their parallel architecture, they are also well adapted for real-time systems due to their fast response and computation times. The training and test data generated by the Min-hash Technique was fed into the DNN algorithm, which produced several hidden layers and generates NN classifiers through training. As compared to alternative works, it has been observed that the proposed method is relatively more effective, as the accuracy rate obtained was remarkably high (98%). The findings show that the signature matrix is more sufficient for this mission, as it emphasizes tempo, secrecy, and honesty. The success criterion for consistency received a high ranking.

ACKNOWLEDGMENT

Authors would like to thank university of Babylon- College of IT for supporting this paper

REFERENCES

[1] D. Coss and S. Samonas, “The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security.,” Journal of Information System Security, vol. 10, no. 3. pp. 21–45, 2014.

[2] T. Sultana, K. A. Sapnaz, F. Sana, and N. Mrs. Jamedar, “email-based-spam-detection.pdf,” Int. J. Eng. Res. Technol., vol. 9, no. 06, June, 2020.

[3] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?,” Proc. 2006 ACM Symp. Information, Comput. Commun. Secur. ASIACCS ’06, vol. 2006, pp. 16–25, 2006, doi: 10.1145/1128817.1128824.

[4] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?,” Proc. 2006 ACM Symp. Information, Comput. Commun. Secur. ASIACCS ’06, vol. 2006, no. March, pp. 16–25, 2006, doi: 10.1145/1128817.1128824.

[5] B. Liang, M. Su, W. You, W. Shi, and G. Yang, “Cracking classifiers for evasion: A case study on the google’s phishing pages filter,” 25th Int. World Wide Web Conf. WWW 2016, pp. 345–356, 2016, doi: 10.1145/2872427.2883060.

[6] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, “Classification using deep learning neural networks for brain tumors,” Future Computing and Informatics Journal, vol. 3, no. 1. pp. 68–71, 2018, doi: 10.1016/j.fcij.2017.12.001.

[7] S. Shikhar and S. Biswas, “Multimodal Spam Classification Using Deep Learning Techniques Shikhar.pdf,” 2017 13th Int. Conf. Signal-Image Technol. Internet-Based Syst., vol. 978-1–5386, pp. 346–349, 2017.

[8] C. C. Aggarwal and C. X. Zhai, “A SURVEY OF TEXT CLASSIFICATION ALGORITHMS,” Elsevier, vol. 9781461432, pp. 163–222, 2013.

[9] F. Wei, H. Qin, S. Ye, and H. Zhao, “Empirical Study of deep learning for text classification in legal document review,” arXiv. 2019.

[10] B. K. Dedeturk and B. Akay, “Spam filtering using a logistic regression model trained by an artificial bee colony algorithm,” Applied Soft Computing Journal, vol. 91. 2020, doi: 10.1016/j.asoc.2020.106229.

[11] M. A. Hassan and N. Mtetwa, “Feature Extraction and Classification of Spam Emails,” 5th International Conference on Soft Computing and Machine Intelligence, ISCMI 2018. pp. 93–98, 2018, doi: 10.1109/ISCMI.2018.8703222.

[12] S. Sumathi and G. K. Pugalendhi, “Cognition based spam mail text analysis using combined approach of deep neural network classifier and random forest,” J. Ambient Intell. Humaniz. Comput., vol. 0123456789, 2020, doi: 10.1007/s12652-020-02087-8.

[13] D. K. Dewangan and P. Gupta, “Email Spam Classification Using Support Vector.pdf,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 6, no. VI, June 2018-Available at www.ijraset.com. 2018.

[14] G. Litjens et al., “A survey on deep learning in medical image analysis,” Med. Image Anal., vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[15] Y. Pan et al., “Brain tumor grading based on Neural Networks and Convolutional Neural Networks,” Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2015-Novem, pp. 699–702, 2015, doi: 10.1109/EMBC.2015.7318458.

[16] “Diving Deep into Deep Learning:History, Evolution, Types and Applications,” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3. pp. 2835–2846, 2020, doi: 10.35940/ijitee.a4865.019320.

[17] A. Anuse and V. Vyas, “A novel training algorithm for convolutional neural network,” Complex Intell. Syst., vol. 2, no. 3, pp. 221–234, 2016, doi: 10.1007/s40747-016-0024-6.

[18] Mohammed Awad and Monir Foqaha, “Email Spam Classification Using Hybrid Approach of Rbf Neural Network and Particle Swarm Optimization,” International Journal of Network Security & Its Applications, vol. 8, no. 5. pp. 19–38, 2016, doi: 10.5121/ijnsa.2016.8402.

[19] A. S. Das, M. Datar, A. Garg, and S. Rajaram, “Google news personalization: Scalable online collaborative filtering,” 16th International World Wide Web Conference, WWW2007. pp. 271–280, 2007, doi: 10.1145/1242572.1242610.

[20] J. Leskovec, A. Rajaraman, and J. D. Ullman, “Mining of Massive Datasets,” Mining of Massive Datasets. 2020, doi: 10.1017/9781108684163.

[21] J. Leskovec, A. Rajaraman, and J. D. Ullman, “Mining of Massive Datasets,” Mining of Massive Datasets. 2014, doi: 10.1017/cbo9781139924801.

[22] M. Majumder, “EMAIL CLASSIFICATION USING ARTIFICIAL NEURAL NETWORK,” pp. 49–54, 2015, doi: 10.1007/978-981-4560-73-3_3.

[23] M. Rout, J. K. Rout, and H. Das, “Correction to: Nature Inspired Computing for Data Science,” vol. SCI 871, 2020, pp. C1–C1.