ISSN 2063-5346

SPAM EMAIL DETECTION USING MACHINE LEARNING

Jagbeer Singh, Satyam Gupta, Shayan Khan, Priyanshu Tyagi, Ketan Chaudhary

Article History: Received: 01.02.2023 Revised: 07.03.2023 Accepted: 10.04.2023

Abstract

Because of the growing use of social media across the globe, the amount of unsolicited huge quantity of e-mails has improved, necessitating the deployment of a truthful machine to filter out such troubles. Junk mail emails are the maximum common trouble at the net. It is straightforward for spammers to ship an e-mail containing spam messages. Spammers have the capability to steal critical records from our gadgets, including files and contacts. Numerous deep studying-based word embedding processes were evolved in recent years. Those tendencies inside the place of phrase illustration can be able to provide a solid strategy to such issues. On this studies, we will observe a method that employs herbal language processing (nlp) to come across junk mail and ham information the use of the junk mail electronic mail dataset. We've got used dense classifier sequential neural communit , lstm and bilstm and in comparison accuracies and consequences. The dataset's efficacy is decided the use of metrics including bear in mind, accuracy, and f1-score.The take a look at indicates that by using bi-lstm category, the dataset's typical accuracy improves. The general paintings is done in python and implemented in a jupyter pocket book

Keywords: *Machine Learning Algorithm, Predictor, airfare, bagging regression*

Meerut Institute of Engineering and Technology, Meerut

DOI:10.31838/ecb/2023.12.s1-B.198

1. INTRODUCTION

E-mail is a form of verbal exchange this is frequently used to talk about a wide variety of information. Due to its convenience and potential to bring messages, files, pix, videos, and hyperlinks, e-mail is broadly utilized with the aid of companies around the sector. Even though the recipient does now not pick out the mail right away, it stays inside the mailbox, geared up to be opened. Furthermore, a single email may be transferred to numerous receivers right away. As a result, email is expedient and price-powerful. Other than the several benefits that emails can provide, unsolicited emails can also occasionally emerge in a user's mailbox. On the internet, spam emails have long been a difficulty. Their excessive volume on the net has a bad effect on the e-mail server's memory. Furthermore, unsolicited mail emails may result in financial loss as a result of malicious customers' deception. Also, unsolicited mail recipients may be inconvenienced and waste time due to the fact they need to examine the entire content to determine whether it is junk mail. They impede the well-timed transmission of valid messages. To hit upon junk mail e-mail is an undertaking for researchers trying to lay out an efficient filtering mechanism due to the aforementioned issues. Information engineering and system-gaining knowledge are famous ways of filtering unsolicited emails. Expertise engineering calls for a hard and fast of rules. It isn't always a powerful mechanism as the rule of thumb set wishes to be continually up to date. Machine studying is proven to be green for this reason as a rule set isn't always required. Gadget gaining knowledge makes use of a set of training and check records. Education records include emails that can be already classified as unsolicited mail or ham. Herbal language processing responsibilities play a crucial position in detecting whether an e-mail is unsolicited or no longer. Nlp converts the unstructured textual content of the email into structured textual content and facilitates text evaluation. Nlp is used to increase the version's accuracy. Numerous strategies in system mastering methodologies may be utilized for e-mail screening.

This paper provides a method for detecting unsolicited mail using herbal language processing techniques. Our aim in this research is to train, check, and examine numerous deep gaining knowledge of and well-known system learning classifiers. The final paper is based on follows-the contribution of researchers in this subject matter is described in phase 2. In section-3,the experimentation framework, dataset, methods, and libraries are described, and in section 4, the precis of the findings are mentioned. Section 5 concludes the paper's scope for destiny..

2 LITERATURESURVEY

Given the importance of the field, some researchers have labored to automate e-mail filtering processes with the use of synthetic intelligence (ai).In this regard [1] proposed an answer for the usefulness of phrase embedding in classifying emails. A pre-skilled transformer version known as Bert (bidirectional encoder representations from transformers) has been quality-tuned to pick out junk mail emails from non-unsolicited mail emails (ham). The consequences are as compared to a dnn (deep neural community) version with two stacked dense layers and a bi-lstm (bidirectional lengthy quick-term reminiscence) layer that was used as a baseline. The consequences are also compared to a fixed of popular classifiers known as ok-nn (ok-nearest neighbors) and nb (nearest neighbor) classifiers (naive Bayes). One open-supply facts set is used to teach the model, and the other is used to get admission to the patience and robustness against unknown statistics. The proposed approach had the very best accuracy of

ninety eight.67% and the very best f1 rating of 98.Sixty six%.

In some other research [2] authors proposed a singular approach based totally on deep getting-to-know (dl) techniques for spam detection on Twitter. To stumble on spammers, the author uses each tweet's content and users' meta-records (i.E. Age of account, variety of followers, and so forth).In [3], the authors consist of a contrast of baseline machine studying techniques for assessment type. To retrieve semantic data, the authors of this paper proposed a technique related to the interest based on bidirectional lstm. In [4]. The authors proposed a unique way of detecting valid messages and nonlegitimate messages. The set of rules used is the horse herd metaheuristic optimization algorithm. The discrete algorithm is constituted of non-stop hoa. The resulting set of rules' inputs was then changed to opposition-primarily based and multi-goal. The proposed method outperforms various other system getting-to-know techniques inclusive of k-nearest neighbors-grey wolf optimization, SVM multilayer perceptron, knn, and naive Bayes, consistent with the assessment findings. In [10], researchers used the lengthy quick-time period memory (lstm) technique to the consciousness of the moods of news tales. Although, they had been constrained by way of useful resource constraints, which intended they did not pay consistent interest to the dependability of news sources themselves.

In [5] authors focused on strategies to efficaciously refine sms junk mail. Numerous device learning-based classifiers just as the naïve bayes, gradient raises logistic regression, and classifier, and deep gaining knowledge of-based totally fashions like cnn and lstm, have been additionally examined. In line with their results, cnn model with the regularisation parameter on randomly generated tenfold move validation facts worked nicely in terms of filtering valid textual content messages, having an accuracy of 99.44%. However, the work changed constrained with the aid of the truth that it relied entirely on messages which might be written in English. In [6] authors uses an included method combining a naive bayes (nb) algorithm and a computational intelligence set of rules which is based totally on particle swarm optimization (PSO). For gaining knowledge of and category of contents of e-mail, the naïve Bayes algorithm is used. As a way to globally optimize the parameters of the naïve Bayes technique, PSO is used. In the evaluation of the man or woman naïve bayes approach, PSO has executed higher performance. In [7] authors, proposed a method that makes use of the wordnet ontology and other methodologies primarily based on semantic and similarity metrics to lower some of the retrieved textual content-based total features, thus lowering the time and area difficulties. Moreover, function choice tactics including the major factor analysis (PCA) and the correlation feature choice are incorporated with dimensionality reduction techniques to gain the smallest ideal characteristic set (cfs). Following the utility of semantic family members, numerous strategies for measuring semantic similarity are used to boost the discount price for the functions. In [8], the authors provided a method for changing the problem of the electronic mail category into a trouble of graph type. This work no longer requires the email text to be transformed into a vector illustration. Whereas this technique converts the e-mail's content into a graph and classifies spam emails through the use of a graph neural community (gun). In [9], authors proposed various methods to locate deceptive news primarily based on information content material utilizing expertise graphs, consisting of the b-transfer mode. Based on the existing transfer version and the new proposed b-transfer model, the creator provided some novel ways for detecting fake information based on incomplete and imperfect understanding graphs

Materials and Methods-This phase describes a method for utilising natural language processing(nlp) to assume unsolicited mail and ham email primarily based at the junk mail e mail dataset. We begin by means of converting, pre-processing, and partitioning datasets to meet the desires of the algorithms underneath consideration. The multiple fashions are then skilled and evaluated, with performance measures used to evaluate and examine them.

3.1 Framework

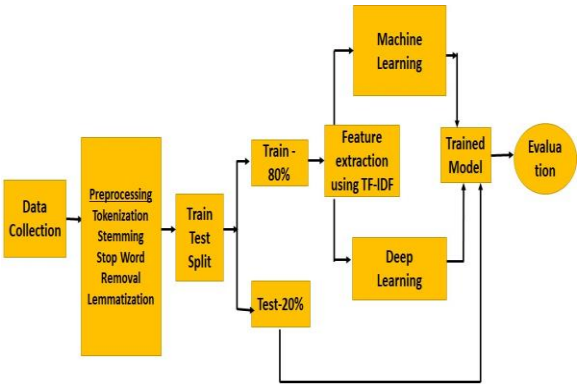

Fig1: Framework of proposed Model

3.1.1 Data Collection - This model uses junk mail e mail dataset from kaggle. The dataset may be viewed at https://www.Kaggle.Com/ datasets/venky73/unsolicited mail-mails-dataset. There are 5171 messages in the dataset. All of the messages are tagged as valid and non-valid.

3.1.2 Data Pre-processing- Pre-processing of dataset is necessary to arrange low-quality statistics into high-quality so that it's miles used similarly. Punctuation symbols, numerals and special characters are eliminated inside the pre-processing step because they don't have any effect on textual content processing. This segment is necessary with a view to put together the text for analysis, modelling, and prediction. Pre-processing includes following natural language processing (nlp) operations inclusive of tokenization of textual content, lowercasing english letters, removal of forestall phrases, stemming and lemmatization. Porter stemmer is used for stemming. Following fig2 suggests the image of information after pre-processing.

	label	message	spam/ham
0	ham	Subject enron methanol meter 988291 follow not...	0
1	ham	Subject hpl nom january 9 2001 see attached fi...	0
2	ham	Subject neon retreat ho ho ho around wonderful...	0
3	spam	Subject photoshop windows office cheap main tr...	1
4	ham	Subject indian springs deal book teco pvr reve...	0

Fig2: Dataset after Pre-processing

The dependencies are imported and junk mail textual content facts are loaded and analysed. Because the records set are unbalanced, we have to use a bar graph to guess the proportion of unsolicited mail and ham. Fig 3a indicates the imbalanced dataset. So that you can solve the problem of imbalanced dataset, we've got used down sampling technique. Some information from the ham class are deleted in order that it fits the count number of minority elegance i.E. Junk mail. Discern 3b suggests the dataset after down sampling

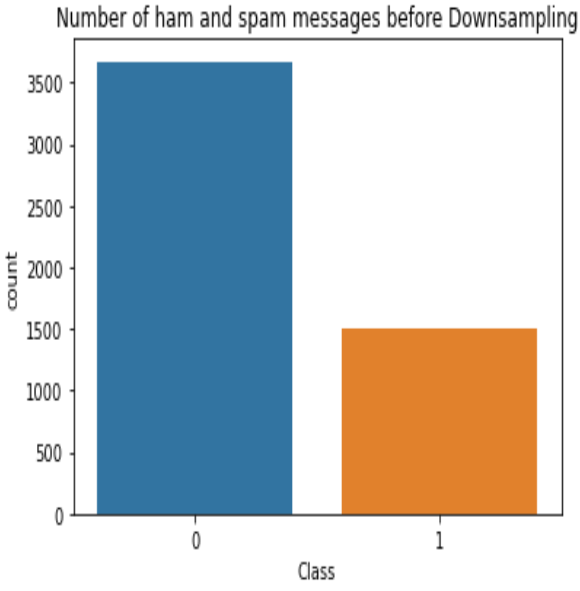

Fig3(a) Unbalanced dataset graph

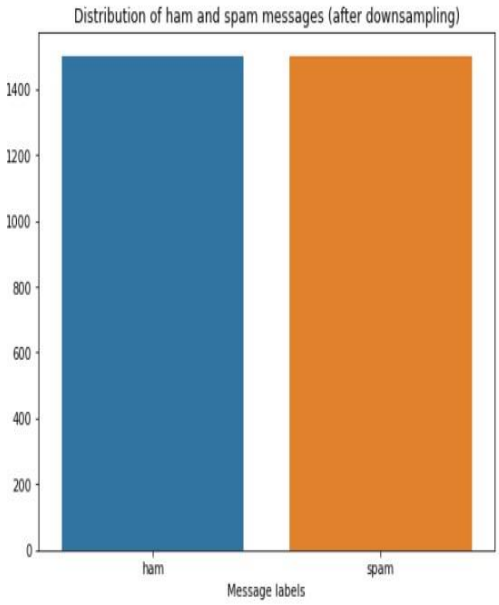

Fig3(b) Balanced dataset graph

i. Because algorithms expect vectors as it is also critical to extract capabilities. Count vectorizer and tokenizer api from tensor waft keras is used for extracting capabilities. Tokenizer api splits sentences into words and performs integer encoding, sequencing represents every sentence into sequences of numbers. A good way to make sequence of identical length, pad sequences() is used. We also divide the records into units: take a look at and training. To govern the functioning of the algorithm, the facts set was separated into eighty% training samples and 20% check samples. Now training of model is achieved the use of deep getting to know strategies and system learning algorithms. The machine studying methods used within the paintings are as follows-

ii. Logistic Regression The cost characteristic of logistic regression is between 0 and 1, and it is a class set of rules based totally on the opportunity concept. The sigmoid feature is used for modelling the records, as visible underneath. We first keep in mind the linear regression characteristic (eqi), then convert the linear function to exponential form as given in eq [11](ii).

$$g(y) = \beta o + \beta\ (Num) \qquad \text{(i)}$$

Here, num is an independent variable, g(y) is the hyperlink function, (p) is the chance of success opportunity is (p) and failure possibility is (1-p). The value of p have to be between 0 and 1, as a result we used this condition to come up with the equation given underneath (ii)$\frac{p}{1-p} = e^{y}$ (ii)

Then, on both sides of the log, we get eq. (iii)

$$\log\left(\frac{p}{1-p}\right) = y$$

Here $\log\left(\frac{p}{1-p}\right)$ is the link

$$\frac{P(y_1\ y_2\ y_3\ y_4 y_5 \ldots, y_n|aj) * P(a_j)}{P(y_1 y_2 y_3 y_4 y_5 \ldots . y_6)}$$

function , y = $\beta o + \beta\ (Num)$, (p/1-p) is the odd ratio.

iii. Naive Bayes- The Bayesian classifier is a probabilistic technique to text categorization that is often used. The main principle of a Bayesian classifier is to determine whether an e-mail is spam or not by examining which words are present in the message and which ones are absent. The Bayesian approach to the new email, according to the literature, is to assign the most likely target label. The simplest version of Bayesian network is a naive Bayes network, in which all characteristics are independent of the class variable's value. The classification problem can be thought of as determining the largest value of the equation below[12].

$$P(aj|y1\ y2\ y3\ y4\ y\ldots,yn) = \text{(iv)}$$

Where $P(a_j)$ is the probability that a random sample falls into category aj .If we already know the training sample is in a_j, $P(y_1, y_2, y_3 \ldots . y_n|a_j)$ is the chance that category $a_j$contains the feature vector

$y = (y_1, y_2, y_3, \ldots, y_n)$.

The joint probability of all possible categories is $P(a_1, a_2, a_3,\ldots..,a_n)$.

`

iv. SVM- Svm stands for assist vector system. It's far a sample category method. Each linearly separable and nonlinearly separable features function well with svm[12]. The general problem is to create a function that decreases mistakes while effectively classifying enter features. The class using the svm algorithm, makes use of the system of making an most reliable line that distinguishes facts factors among two lessons. The overall performance of classifiers for small sample gaining knowledge of troubles has been more desirable by means of the usage of sequential minimal optimization (smo) and a polynomial kernel function. The subsequent is the definition of the svm choice feature:

$$F(z) = \sum_{i=1}^{n} \beta K(y_i, z) + b \quad \text{(vi)}$$

$K(y_i, z)$ is the kernel function. The unclassified tested feature is Z. The support vectors are y_i and their weights are β and the constant bias is b.

v. Random Forest: It is a approach of machine learning used for solving classification & regression issues. Random Forest makes use of ensemble learning to solve complex problems by combining many classifiers.

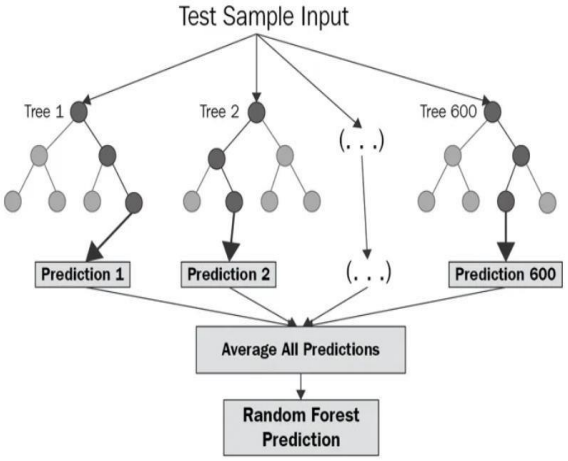

Fig. 4.:Random forest architecture [13]

It accommodates of a big number of selection timber, often referred to as estimators, as proven in determine 4. Every node within the tree is taught to make its predictions using a separate set of observations [13]. The random woodland's very last predictions are then calculated through averaging the forecasts of every tree.It is solved by eqs. (vii-x)

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \quad \text{(vii)}$$

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \quad \text{(viii)}$$

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \quad \text{(ix)}$$

$$ni_j = W_j C_j - W_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)}$$

v. Decision Tree-Any other set of rules that has been hired greater regularly inside the supervised gaining knowledge of method studies is the choice tree gadget mastering set of rules. It has the capacity to work with both numerical and categorical statistics in a system. The output of the selection tree classifier is comparable to that of a binary tree, often referred to as a choice tree. A tree is made of branches, each of which represents a set of alternatives to pick from, in addition to leaf nodes, that are used for classification. It makes use of associations guidelines for predicting and associating target labels[14].

vi. .LSTM-In text evaluation, for semantic composition, long brief term memory is the maximum latest deep gaining knowledge of approach. Features derived from sentence illustration are used to predict the score of critiques. In phrase embedding, every phrase within the textual content corpus is shown as a word vector with a non-stop, low-dimensional, and actual fee. The phrase vectors are located, wherevshows the vocabulary size and d denotes the word dimension. The sentence vectors were computed using lstm from the word vector. Lstm is anrnn version that iteratively transforms the vector of phrase wt to the output phrase vector of its preceding step ht-1 to switch various length

phrase vectors to constant-duration word vectors. The standard rnn, on the other hand, suffers from the hassle of vanishing gradient, wherein gradients construct or decay exponentially over extended sequences. While the internal operating competencies of rnn and lstm cells are as compared, it's far discovered that the long quick term memory cellular has 3 extra gates: an input gate, output gate, and forget gate. After forgetting previous history, these gates don't forget the value input and output vector . The lstm mobile's computation is as follows [16]:

$ig_t = \sigma\ (W_{ig}\ .\ [h_{t-1}, y_t] + b_{ig}\ ])$

$fg_t = \sigma(W_{fg}\ .\ [h_{t-1}, y_t\ ] + b_{fg}\ ])$ (xii) $o_{gt} = \sigma\ )$

$(W_{og}\ .\ [h_{t-1}, y_t] + b_{og}\ ])$ (xiii)

$\check{C}g_t = tanh(W_{cg}\ .\ [h_{t-1}, y_t] + b_{cg})$ (xiv)

$Cgt = fgt\ \ Ct\text{-}1 + igt\check{C}gt$ (xv) $h_t = o_{gt} tanh(Cg_t)$ (xvi)

W_{ig} ,W_{og} ,W_{fg} , $b_{ig,}$ $b_{og,\ and}$ $b_{fg}$are input gate, output gate and forget gate respectively. The Long short term memory model considers the last hidden vector as the sentence representation after calculating the hidden vector for each position,

vii. Bi-LSTM-Only the preceding context can be accessed by the unidirectional LSTM**.** However, most of the sequential data, particularly in the classification problems, are based on previous states and next states. As a result, we chose bidirectional architecture to process the sequencing in both directions i.e. left to right which is called Forward LSTM and right to left which is called Backward LSTM, resulting in getting overall sequence information .The Bi-LSTM layer also receives the input obtained from the Dropout layer.

4. Experimental Results and Analysis-- in this segment, we present the effects of the experimental assessment carried out to reveal the effectiveness of the proposed technique. We have used enron dataset for evaluating the outcomes. Dataset is obtained from kaggle.Com. Python 3.7 is utilized to run python code for the version's implementation. The model is educated the use of a couple of classifiers to test and evaluate the consequences for higher accuracy. Each classifier gives its evaluated results to the user. Five exceptional system learning algorithms had been implemented. With the intention to carry out a comparative analysis in their performance, various key overall performance metrics which includes accuracy, precision, take into account and f1 rating are considered. Desk 1 indicates the findings of gadget getting to know classifier on checking out data. Random wooded area has executed (xi) maximum accuracy of ninety six.Eight% amongst all device gaining knowledge of classifiers the dense sequential junk mail detection version, long brief time period reminiscence (lstm) version and bi-lstm is carried out using the keras python api. For binary multi-label type duties, the keras deep getting to know package makes it easy to define and compare neural network models. The findings of deep studying classifiers on testing records are supplied in desk ii. Deep learning architectures have made widespread upgrades over traditional device getting to know strategies, consistent with the findings. In deep gaining knowledge of strategies, bi-lstm has completed maximum validation accuracy of ninety nine%. Bi-lstm may be taken into consideration as first-class model for electronic mail spam detection does the sequencing in each the directions i.E. Preceding states and the next states.

Table1: Machine Learning Classifier Results evaluated on testing data

		Precision	Recall	F1-Measure	Accuracy
Logistic Regression	Not spam	01	.93	.96	95.86
	Yes Spam	0.91	1	.951	
Naive Bayes	Not spam	1.00	.94	.972	96.3
	Yes Spam	.92	1.00	0.96	
Random Forest	Not spam	0.99	.95	0.97	96.85
	Yes Spam	.94	.99	.96	
DT	Not span	.961	.91	.93	92.6
	Yes Spam	.895	.94	.92	
Gradient Boosting	Not spam	.99	.92	.96	95
	Yes Spam	.90	.99	.94	
Dense Sequential Model	Not spam	.985	.94	.96	97
	Yes Spam	.94	.98	.96	
LSTM	Not spam	0.98	.92	.95	98.3
	Yes Spam	0.93	.98	.95	
Bi-LSTM	Not spam	0.98	.94	.96	98.5
	Spam	0.94	.94	.96	

Table2: Deep Learning Classifier Results evaluated on testing data

	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Dense Sequential Model	0.04	0.98	0.11	0.959
LSTM	0.06	0.97	0.12	0.958
Bi- LSTM	0.03	0.99	0.18	0.956

6 CONCLUSION AND FUTURESCOPE

This paper discusses a number of strategies that may be used to categorise spam and ham electronic mail. Experiments are performed using various gadget and deep getting to know classifiers. Effects show that bilstm has most accuracy of ninety eight.5percentand f1-measure 0f 96%. Within the future, the present paintings have to be stepped forward by way of extending it to an expansion of fields, which include e-commerce, job profile-based totally web sites, and different locations where fake information is conventional, as well as an software that allows users to detect faux facts using their cellphone in a shorter quantity of time. Also, paintings may be prolonged via implementing the classifier in real time

7 REFERENCES

[1] AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," in *Procedia Computer Science*, 2021, vol. 184, 853–858 doi: 10.1016/j.procs.2021.03.107.

[2] S. Madisetty and M. S. Desarkar, "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 4, pp. 973–984, Dec. 2018, doi: 10.1109/TCSS.2018.2878852.

[3] A. Salunkhe, "Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.14789

[4] A. Hosseinalipour and R. Ghanbarzadeh, "A novel approach for spam detection using horse herd optimization algorithm," *Neural Comput. Appl.*, Mar. 2022, doi: 10.1007/s00521-02207148-x.

[5] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Futur. Gener. Comput. Syst.*, vol. 102, pp. 524–533, Jan. 2020, doi: 10.1016/j.future.2019.09.001.

[6] E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, "Efficient email classification approach based on semantic methods," *Ain Shams Eng. J.*, vol. 9,3259–3269, 2018, doi: 10.1016/j.asej.2018.06.001.

[7] W. Pan *et al.*, "Semantic Graph Neural Network: A Conversion from Spam Email Classification to Graph Classification," *Sci. Program.*, vol. 2022, 2022, doi: 10.1155/2022/6737080.

[8] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," in Lecture Notes in Computer Science (including

subseries Lecture Notes in AI and Lecture Notes in Bioinformatics, 2018, vol. 11136, pp. 669–683. doi: 10.1007/978-3-030-00671-6_39.

[9] Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. Transactions on Emerging Telecommunications Tech, 31(2), e3767.

[10] W. Feng, "2016 IEEE 35th International Performance Computing and Communications Conference, IPCCC 2016," *2016 IEEE 35th Int.*