FAKE NEWS DETECTION ON KAGGLE DATASETS USING DECISION TREE ALGORITHM

Mrs.P. Manjula M.Tech,Phd
Computer Science and Business Systems
Panimalar Engineering College
Chennai, India

G. Priyanka
Computer Science and Business Systems
Panimalar Engineering College
Chennai, India

N. Nandhini Sai
Computer Science and Business Systems
Panimalar Engineering College
Chennai, India

Poojakaran D
Computer Science and Business Systems
Panimalar Engineering College
Chennai, India

Subalakshmi R
Computer Science and Business Systems
Panimalar Engineering College
Chennai, India

Sneha R
Computer Science and Business Systems
Panimalar Engineering College
Chennai, India

ABSTRACT

Today, social media is widely used to disseminate real-time news. Users of social media are scattered over a wide range of people and do not belong to any one age or gender category. This is due to the fact that it is simple to transmit, spreads quickly, is simple to access, and costs little to spread the word. Social media users are utilising it to disseminate fake news, and malicious information for purposes of commerce, politics, and entertainment. This dissertation offered four approaches to use in judging the veracity of the news, so limiting its dissemination, to address these challenges. The sharing of information via the internet has been increasing over the years. The internet has been a source of easy information and is used more than traditional ways like newspapers or magazines. It is important to identify information from the internet as real or fake as misleading information could cause a lot of havoc in society. Fake information can be the cause of riots, and chaos and can affect a large group of society. In this paper,the methodology used to detect false news using machine learning classifiers to authenticate whether the news is real or not. For the generation of feature vectors, TF-IDF Vectorizer have been utilized. To detect the news as fake or real ,the proposed approach is compared with several

Keywords: - Fake news detection, logistic regression, decision tree, random forest, gradient boosting, TF-IDF Vectorizer, Machine learning.

1.INTRODUCTION

After the middle of the 1990s, the World Wide Web underwent a significant development, and online social media began to be used for interpersonal communication. The majority of people use social media platforms to communicate opinions, and they come from a variety of ages, genders, and communities, according to research by Jiang et al. [1]. Users

utilise Twitter and Facebook, two software behemoths in the social media space, to share real-time news. Due to their rapid information sharing, low cost, and ease of use, social media platforms are now regarded as key platforms.Videos, photos, or text that is sent to spread false information with false facts is referred to as fake news. Even though the news may appear to be true at first, it will elicit surprising responses and draw readers'attention. They are made by organisations or people who are driven by their own interests, which could be driven by personal, political, or economic agendas. With the advent of print media, bogus news

has been widely disseminated. But because of the digital exchange, they are becoming prevalent and common. Fake news is more visible in a shorter period of time because social media sharing is simple and quick.

2. RELATED WORKS

The main focus of Georgios et alstudy .'s article [5] is the identification of bogus news. The author chose machine learning methods to solve the problem and employed the content-based feature to produce accurate findings. With the help of the author's experiments, it was simple to identify between phoney and authentic news. The investigation began with a thorough feature analysis of the data. The author decided to use a variety of machine learning techniques and ensemble algorithms, which are quite effective at handling text classification tasks, to detect fake news in written narratives and word embeddings. AdaBoost, SVM, DT, Bagging, and other algorithms were utilised by the author, who also decided to undertake tests using obsolete data sources. The goal of the study work by Cody et al. [3] is to automatically detect false information in Twitter discussions. The author devised the automatic detection method after looking into the CREDBANK and PHEME accuracy assessment processes. The author conducted experiments using three datasets: PHEME, a prospective rumours dataset with journalists' accuracy assessments; CREDBANK, a crowdsourced dataset of accuracy evaluation; and BuzzFeed's false news dataset. The author initially found certain features falling under four kinds, aligned the three datasets in a consistent way, employed classifiers, and then analysed each feature set under the receiver operating characteristic (ROC)curve to construct the model for detection.

Judee et al. [8] idea of making use of linguistic traits as instruments to identify dishonesty in communication established the groundwork for this study paper. The author first chose two tests for the research: surviving in the desert and flying. Nonetheless, the outcomes of these two tests were inconsistent. As a result, the author chose to conduct in-depth research on a theft scenario. Results were analysed using both individual and cluster cues. Data mining algorithms were employed for cluster analysis, which assisted the author in creating an automatic criteria for spotting fraud. The author used C4.5 [1], which eliminated unnecessary branches and limited error rates.

According to Zubair et al. [6], misleading news can be filtered or classified using machine learning techniques including SVM, RF, NB, and DT. For this, the author selected a few sets of both fake and real news stories. On the basis of the texts of the news stories he has chosen, the author attempts to construct a classification strategy. The aforementioned classifiers were applied to AdaBoost and Bagging during the classification process. With the help of Pycharm in the Python environment, these tests were run. Classification parameters like recall, ROC, F-score, accuracy, and precision were used to gauge the classifier's performance. All of these examinations of the news have taught the author how to identify fake news and locate its source. The top-performing classifiers were determined by comparing the classification metrics ofseveralclassifiers. In order to combat fake news, which poses a serious threat to several sectors, Rohit et al. [10] propose a deep convolutional neural network called the Futue Network Development. Instead of depending on manually created characteristics, the author would like to create a model that uses numerous hidden layers in a deep neural network to automatically learn various discriminatory features for the categorization of fake news. At each layer, the CNN will assist in extracting features whose performance may be evaluatedagainstbenchmarkmodels. Tao Jiang [11]hold-out cross-validation was used to evaluate how well three deep learning models and five machine learning models performed on two fake and real news datasets of varied sizes. The models' performance was evaluated using accuracy, precision, recall, and F1-score, and a modified version of McNemar's test was used to determine whether there was a significant difference. Then, we showed our unique stacking model, which, when applied to the ISOT dataset and the KDnugget dataset, respectively, achieve, a testing accuracy of 99.94% and 96.05%. In addition, our proposed approach outperforms conventional methods.

3. PROPOSED SYSTEM DESIGN

The proposed system classifies new articles as true or false based on the currently available data when presented with a scenario of a group of news items. This forecast is based on how the terms used in the article relate to one another. By employing a Word2Vec model to identify the links between words, the suggested method can categorize new articles into those that are true and those that are false.

Datasets are used to hold input that is gathered from a variety of sources, including newspapers and social media. Datasets will be used as input by the system. The datasets go through preprocessing, when the extraneous data is deleted and, if necessary, the data types of the columns are altered. In the preceding phase, Google Collab is utilized. Dataset is used to train the machine for fake news detection. The entire dataset is split into two datasets before commencing the false news identification phase. 20% is used for testing, while the remaining 80% is for training. The model is trained using the training dataset by machine learning algorithms. The test dataset is used as the input, and a projected result is produced. To determine the number of accurate and incorrect predictions in the context of true and false news, the predicted and actual outputs are compared.

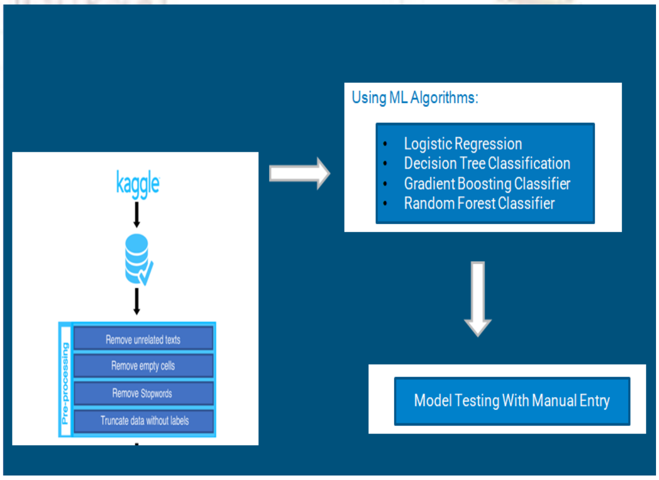


FIGURE1- PROPOSED SYSTEM DESIGN

4. IMPLEMENTATION

Importing Required Library

from google.colab import drive
drive.mount('/content/drive') import pandas as pd
import numpy as np import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report import
import string

Inserting Fake And Real DataSet

df_fake=

pd.read_csv("/content/drive/MyDrive/Fake.csv.zip")

df_true=

pd.read_csv("/content/drive/MyDrive/True.csv.zip")

df_fake.head(5)

df_true.head(5)

Inserting a column called "class" for fake and real news dataset to categoriesfake and true news.

df_fake["class"] = 0

df_true["class"] = 1

Removing last 10 rows from both the dataset, for manual testing

df_fake.shape, df_true.shape df_fake_manual_testing = df_fake.tail(10) for i inrange(23480,23470,-1):

df_fake.drop([i], axis = 0, inplace = True) df_true_manual_testing = df_true.tail(10) for i inrange(21416,21406,-1):

df_true.drop([i], axis = 0, inplace = True) df_fake.shape, df_true.shape

Merging the manual testing dataframe in single dataset and save it in a CSV file

df_fake_manual_testing["class"] = 0

df_true_manual_testing["class"]=1 df_fake_manual_testing.head(10) df_true_manual_testing.head(10)

Merging the main fake and true dataframe

df_marge = pd.concat([df_fake, df_true], axis =0) df_marge.head(10)

df_marge.columns

5.PERFORMANCE METRICS

Performance metrics are important in evaluating the effectiveness of any machine learning algorithm, including decision tree algorithms used for fake news detection. The following are some performance metrics that can be used in evaluating the effectiveness of decision tree algorithm for fake news detection:

Accuracy

This measures the proportion of correct classifications to total classifications made by the decision tree model. A higher accuracy indicates that the model is performing well in distinguishing between fake and real news.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \longrightarrow (1)$$

Precision

This measures the proportion of true positives to the total number of positive classifications. A higher precision indicates that the model is able to accurately identify fake news.

$$\text{Precision} = \frac{TP}{TP + FP}. \longrightarrow (2)$$

Recall

This measures the proportion of true positives to the total number of actual positives. A higher recall indicates that the model is able to identify more of the fake news present in the dataset.

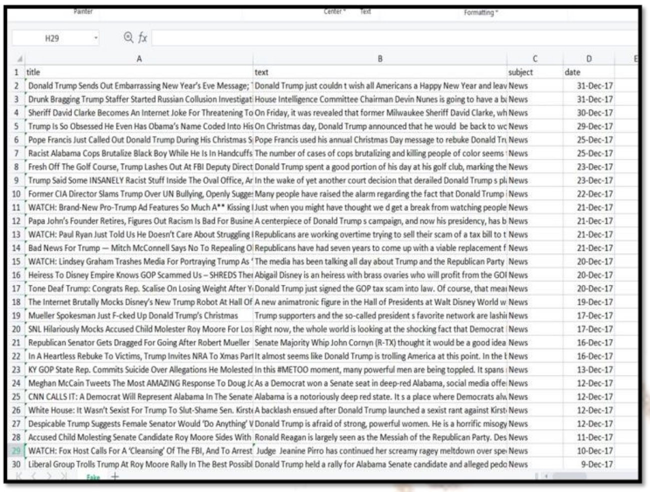
$$\text{Recall} = \frac{TP}{TP + FN} \longrightarrow (3)$$

F1-Score

This is the harmonic mean of precision and recall, and provides a combined metric of the two. A higher F1 score indicates that the model is able to achieve high precision and recall

$$\text{F1} - \text{score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \longrightarrow (4)$$

6.SCREENSHOTS



	title	text	subject	date
2	Donald Trump Sends Out Embarrassing New Year's Eve Message;	Donald Trump just couldn t wish all Americans a Happy New Year and leav	News	31-Dec-17
3	Drunk Bragging Trump Staffer Started Russian Collusion Investigat	House Intelligence Committee Chairman Devin Nunes is going to have a b	News	31-Dec-17
4	Sheriff David Clarke Becomes An Internet Joke For Threatening To	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, wh	News	30-Dec-17
5	Trump Is So Obsessed He Even Has Obama's Name Coded Into His	On Christmas day, Donald Trump announced that he would be back to wc	News	29-Dec-17
6	Pope Francis Just Called Out Donald Trump During His Christmas S	Pope Francis used his annual Christmas Day message to rebuke Donald Tr	News	25-Dec-17
7	Racist Alabama Cops Brutalize Black Boy While He Is In Handcuffs	The number of cases of cops brutalizing and killing people of color seems	News	25-Dec-17
8	Fresh Off The Golf Course, Trump Lashes Out At FBI Deputy Direct	Donald Trump spent a good portion of his day at his golf club, marking the	News	23-Dec-17
9	Trump Said Some INSANELY Racist Stuff Inside The Oval Office, Ar	In the wake of yet another court decision that derailed Donald Trump s pl	News	23-Dec-17
10	Former CIA Director Slams Trump Over UN Bullying, Openly Sugge	Many people have raised the alarm regarding the fact that Donald Trump	News	22-Dec-17
11	WATCH: Brand-New Pro-Trump Ad Features So Much A** Kissing I	Just when you might have thought we d get a break from watching people	News	21-Dec-17
12	Papa John's Founder Retires, Figures Out Racism Is Bad For Busine	A centerpiece of Donald Trump s campaign, and now his presidency, has b	News	21-Dec-17
13	WATCH: Paul Ryan Just Told Us He Doesn't Care About Struggling I	Republicans are working overtime trying to sell their scam of a tax bill to t	News	21-Dec-17
14	Bad News For Trump — Mitch McConnell Says No To Repealing O	Republicans have had seven years to come up with a viable replacement f	News	21-Dec-17
15	WATCH: Lindsey Graham Trashes Media For Portraying Trump As '	The media has been talking all day about Trump and the Republican Party	News	20-Dec-17
16	Heiress To Disney Empire Knows GOP Scammed Us – SHREDS Thei	Abigail Disney is an heiress with brass ovaries who will profit from the GOI	News	20-Dec-17
17	Tone Deaf Trump: Congrats Rep. Scalise On Losing Weight After Y	Donald Trump just signed the GOP tax scam into law. Of course, that mea	News	20-Dec-17
18	The Internet Brutally Mocks Disney's New Trump Robot At Hall Of	A new animatronic figure in the Hall of Presidents at Walt Disney World w	News	19-Dec-17
19	Mueller Spokesman Just F-cked Up Donald Trump's Christmas	Trump supporters and the so-called president s favorite network are lashi	News	17-Dec-17
20	SNL Hilariously Mocks Accused Child Molester Roy Moore For Los	Right now, the whole world is looking at the shocking fact that Democrat	News	17-Dec-17
21	Republican Senator Gets Dragged For Going After Robert Mueller	Senate Majority Whip John Cornyn (R-TX) thought it would be a good idea	News	16-Dec-17
22	In A Heartless Rebuke To Victims, Trump Invites NRA To Xmas Part	It almost seems like Donald Trump is trolling America at this point. In the	News	16-Dec-17
23	KY GOP State Rep. Commits Suicide Over Allegations He Molested	In this #METOO moment, many powerful men are being toppled. It spans	News	13-Dec-17
24	Meghan McCain Tweets The Most AMAZING Response To Doug Jc	As a Democrat won a Senate seat in deep-red Alabama, social media offe	News	12-Dec-17
25	CNN CALLS IT: A Democrat Will Represent Alabama In The Senate	Alabama is a notoriously deep red state. It s a place where Democrats alw	News	12-Dec-17
26	White House: It Wasn't Sexist For Trump To Slut-Shame Sen. Kirst	A backlash ensued after Donald Trump launched a sexist rant against Kirst	News	12-Dec-17
27	Despicable Trump Suggests Female Senator Would 'Do Anything' V	Donald Trump is afraid of strong, powerful women. He is a horrific misogy	News	12-Dec-17
28	Accused Child Molesting Senate Candidate Roy Moore Sides With	Ronald Reagan is largely seen as the Messiah of the Republican Party. Des	News	11-Dec-17
29	WATCH: Fox Host Calls For A 'Cleansing' Of The FBI, And To Arrest	Judge Jeanine Pirro has continued her screamy ragey meltdown over spe	News	10-Dec-17
30	Liberal Group Trolls Trump At Roy Moore Rally In The Best Possibl	Donald Trump held a rally for Alabama Senate candidate and alleged pedo	News	9-Dec-17

FIGURE 6.1-DATA COLLECTION

News from six distinct industries are included in the Fake News Database dataset: technology, business, education, sports, politics, and entertainment. The dataset's real news was gathered from numerous popular news websites, primarily in the US, including ABC News, CNN, USA Today, New York Times, Fox News, Bloomberg, and CNET, among others.

1. Logistic Regression

```
[28] from sklearn.linear_model import LogisticRegression

[29] LR = LogisticRegression()
     LR.fit(xv_train,y_train)

     LogisticRegression()

[30] pred_lr=LR.predict(xv_test)

[31] LR.score(xv_test, y_test)

     0.9862745098039216

[32] print(classification_report(y_test, pred_lr))

                  precision    recall  f1-score   support

               0       0.99      0.98      0.99      5849
               1       0.98      0.99      0.99      5371

        accuracy                           0.99     11220
       macro avg       0.99      0.99      0.99     11220
    weighted avg       0.99      0.99      0.99     11220
```

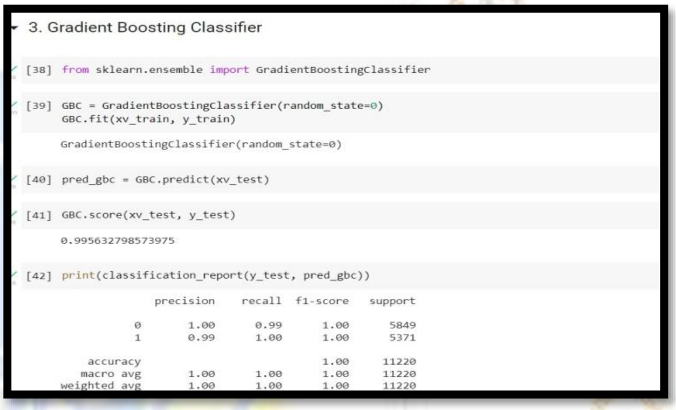
If the outcome variable only has two possible values, the model is referred to as a binary logistic regression model.

Both real news (Y = 1) and fraudulent news (Y = 0) are possible in this situation. The chance that a record belongs to a positive class when Y = 1 is represented by P (Y=1) = ez/ (1 + (ez)) in the binary logistic regression model.Based on previous data set observations, the logistic regression statistical analysis technique can be used to predict a binary result, such as yes or no (binary classification). A supervised statistical method is used to determine the likelihood of the dependent variable.

2. Decision Tree Classification

```
[33] from sklearn.tree import DecisionTreeClassifier

[34] DT = DecisionTreeClassifier()
     DT.fit(xv_train, y_train)

     DecisionTreeClassifier()

[35] pred_dt = DT.predict(xv_test)

[36] DT.score(xv_test, y_test)

     0.996078431372549

[37] print(classification_report(y_test, pred_dt))

                  precision    recall  f1-score   support

               0       1.00      1.00      1.00      5849
               1       1.00      0.99      1.00      5371

        accuracy                           1.00     11220
       macro avg       1.00      1.00      1.00     11220
    weighted avg       1.00      1.00      1.00     11220
```

Combining gradient descent and boost is known as gradient boosting. For each successive model in gradient boosting, the loss function from the preceding model is scaled down using the gradient descent method. By repeating this process, the estimation of the target variable is improved.



3. Gradient Boosting Classifier

```
[38] from sklearn.ensemble import GradientBoostingClassifier

[39] GBC = GradientBoostingClassifier(random_state=0)
     GBC.fit(xv_train, y_train)

     GradientBoostingClassifier(random_state=0)

[40] pred_gbc = GBC.predict(xv_test)

[41] GBC.score(xv_test, y_test)

     0.995632798573975

[42] print(classification_report(y_test, pred_gbc))

                  precision    recall  f1-score   support

               0       1.00      0.99      1.00      5849
               1       0.99      1.00      1.00      5371

        accuracy                           1.00     11220
       macro avg       1.00      1.00      1.00     11220
    weighted avg       1.00      1.00      1.00     11220
```

4. Random Forest Classifier

```
[43] from sklearn.ensemble import RandomForestClassifier

[44] RFC = RandomForestClassifier(random_state=0)
     RFC.fit(xv_train, y_train)

     RandomForestClassifier(random_state=0)

[45] pred_rfc = RFC.predict(xv_test)

[46] RFC.score(xv_test, y_test)

     0.9905525846702318

[47] print(classification_report(y_test, pred_rfc))

                  precision    recall  f1-score   support

               0       0.99      0.99      0.99      5849
               1       0.99      0.99      0.99      5371

        accuracy                           0.99     11220
       macro avg       0.99      0.99      0.99     11220
    weighted avg       0.99      0.99      0.99     11220
```

A method that lowers the variance of an estimated function of prediction is known as bagging or bootstrap aggregation. Bagging functions effectively with high variance and low-bias categorization algorithms like trees. As a notable advancement in bagging, random forests represent a considerable group. Take an average for the associated trees after that. With no increase in variance, Random Forest improved on bagging by reducing correlation across trees. Since they are easier to train and tune, random forest performance often resembles that of boosting.

Random forests are hence popular algorithms that are used with different packages

7. CONCLUSION AND FUTURE WORK

As a result, the model uses modules to deal directly with cleaned data. Additionally, algorithms were used to train the data. Manually classifying news demands in-depth subject knowledge and the ability to spot irregularities in the content. Using ensemble methods and machine learning models, we examined the problem of categorising bogus news items in this study. Instead of expressly classifying political news, the data we used in our study was compiled from news articles from a number of domains that cover the majority of e-news. The basic goal of the search is to identify textual patterns that discriminate between false and legitimate news. The learning models were trained and parameterized to attain the highest level of accuracy. Compared to other models, some have attained a better level of accuracy. To evaluate the effectiveness of each method, we used a variety of performance indicators. Comparing the resemble learners to the individual students, the resemble learners have often performed better. Researchers need to focus on several open problems in fake news identification. Machine learning algorithms can be used to identify the primary sources engaged in the dissemination of fake news, for example, to decrease the spread of fake news, identifying key factors involved in the spread of news is an important first step .Put the classifiers together to improve performance. Check the news's sources. Search the news online to examine the news content, on the one hand, warning news consumers and promoting tools so they can be informed and question the sources of information is a very positive thing, but on the other hand, we might be creating news consumers who don't believe in the value of well-sourced news and distrust everything. The latter course could lead to a general condition of confusion where news consumers are indifferent or unable to judge the reliability of any news source.

8. REFERENCES

[1] O. Ngada and B. Haskins, "Fake News Detection Using Content-Based Features and Machine Learning," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi:10.1109/CSDE50874.2020.9411638.

[2] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900-903, doi:10.1109/UKRCON.2017.8100379.

[3] N. F. Baarir and A. Djeffal, "Fake News detection Using Machine Learning," 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well- being (IHSH), 2021, pp. 125-130, doi:10.1109/IHSH51661.2021.9378748.

[4] N. S. Yuslee and N. A. S. Abdullah, "Fake News Detection using Naive Bayes ," 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), 2021, pp. 112-117, doi:10.1109/ICSET53708.2021.9612540.

[5] J. C. S. Reis, A. Correia, F. Murai, A. Veloso and F. Benevenuto, "Supervised Learning for Fake News Detection," in IEEE Intelligent Systems, vol. 34, no. 2, pp. 76- 81, March-April 2019, doi:10.1109/MIS.2019.2899143.

[6] J. Shaikh and R. Patil, "Fake News Detection using Machine Learning," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), 2020, pp. 1-5, doi:10.1109/iSSSC50941.2020.9358890.

[7] J. Zhang, B. Dong and P. S. Yu, "Fake Detector: Effective Fake News Detection with Deep Diffusive Neural Network," 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020, pp. 1826-1829, doi:10.1109/ICDE48307.2020.00180.

[8] G. Anusha, G. Praveen, D. Mounika, U. S. Krishna and R. Cristin, "Detection of Fake News using machine learning," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 2022, pp. 1-5, doi: 10.1109/ICDCECE53908.2022.9793155.

[9] S. Lyu and D. C. -T. Lo, "Fake News Detection by Decision Tree," 2020 SoutheastCon, 2020, pp. 1-2, doi:10.1109/SoutheastCon44009.2020.9249688.

[10] Jain and A. Kasbe, "Fake News Detection," 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018, pp. 1-5, doi:10.1109/SCEECS.2018.8546944.

[11] A.Douglas,"News consumption and then electronic media," 7e International Journal of Press/Politics, vol.11,no.1,pp.29–52,2006.

[12]J.Wong,"Almostallthetrafficto fakenewssitesisfromfacebook,newdatashow,"2016.

[13]S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: "Evidence from Financial Markets,"2019.

[14] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "ex BAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert)." Applied Sciences, vol. 9, no. 19, 2019.

[15] S. Akhtar, F. Hussain, F. R. Raja et al., "Improving mispronunciation detection of arabic words for non-native learners using deep convolutional neural network features," Electronics, vol. 9, no. 6,2020.

[16] Survey on Fake News Detection using Machine learning Algorithms,ICACT-2021.

[17] Fake News Classification Using Random Forest and Decision Tree, December,2020.
[18] Priyanka S, "Detection of Fake Profiles on Twitter using Random Forest &Deep Convolutional Neural Network",2019.