# COMPARATIVE STUDY ON FAKE NEWS DETECTION

**A PROJECT REPORT**

*for*

**NATURAL LANGUAGE PROCESSING (SWE1017)**

*in*

**M.Tech (Software Engineering)**

*by*


**NITHISH KUMAR S (20MIS0024)**

**SHASHIKIRAN L (20MIS0274)**

**DINESH RAJAN S (20MIS0449)**


**7th Semester, 4th Year**

*Under the Guidance of*

**Prof. SENTHILKUMAR M**

Associate Professor Sr, SITE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

NOV, 2023

<u>**DECLARATION BY THE CANDIDATE**</u>

We hereby declare that the project report entitled **"COMPARATIVE STUDY ON FAKE NEWS DETECTION"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Natural Language Processing (SWE1017)** is a record of bonafide project work carried out by us under the guidance of **Prof. Senthilkumar M.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

**NITHISH KUMAR S**

Place: Vellore

Signature

Date: 04/11/2023

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

## CERTIFICATE

This is to certify that the project report entitled **"COMPARATIVE STUDY ON FAKE NEWS DETECTION"** submitted by **Nithish Kumar S (20MIS0024), SHASHIKIRAN L (20MIS0274), DINESH RAJAN S (20MIS0449)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Natural Language Processing (SWE1017)** is a record of bonafide work carried out by them under my guidance.

**Prof. Senthilkumar M**

**GUIDE**

**Asso. Professor Sr, SITE**

# ABSTRACT

Fake news has become a major problem in the digital age, with the potential to mislead and misinform the public. Logistic regression is a machine learning algorithm that can be used to classify data, making it a promising tool for fake news detection. In this paper, we propose a novel fake news detection model using logistic regression. Our model is based on a set of features extracted from news articles, including lexical features, syntactic features, and semantic features. We evaluate our model on a benchmark fake news dataset and achieve an accuracy of 100% on the training data and 100% on the test data. Our results show that logistic regression is a promising algorithm for fake news detection. Our model is simple to implement and can be trained on relatively small datasets. Additionally, our model is able to achieve high accuracy on a benchmark dataset, suggesting that it can be used to effectively detect fake news in real-world applications. Fake news has become a major problem in recent years, as it can have a significant impact on public opinion and decision-making. Machine learning has been proposed as a promising approach to detecting fake news, as it can be used to analyze large amounts of data and identify patterns that may indicate the presence of false or misleading information. A number of recent research papers have explored the use of machine learning for fake news detection. For example, in [1], the authors propose a model that uses a combination of features, including the text of the article, the source of the article, and the social media engagement of the article, to predict whether the article is fake or real. The model achieves an accuracy of 90.46% on a dataset of fake and real news articles. In [2], the authors propose a model that uses deep learning to detect fake news. The model is trained on a dataset of fake and real news articles, and it learns to extract features from the text of the articles that are indicative of whether the article is fake or real. The model achieves an accuracy of 92.5% on the dataset.

# INTRODUCTION

Fake news, or misinformation, is a serious problem that has the potential to undermine democracy, public trust, and social cohesion. In recent years, there has been a growing interest in developing machine learning models to detect fake news. One promising approach is to use logistic regression, which is a simple but effective classification algorithm. Logistic regression is a supervised learning algorithm that can be used to predict the probability of a binary outcome, such as whether a news article is fake or real. The algorithm works by learning a linear relationship between a set of input features and the target output variable. In the context of fake news detection, the input features could be things like the title of the article, the body of the article, the source of the article, and the social media engagement of the article. Logistic regression has several advantages over other machine learning algorithms for fake news detection. First, it is relatively simple to understand and implement. Second, it is interpretable, meaning that it is possible to understand why the algorithm makes the predictions that it does. Third, it is robust to overfitting, which is a common problem with machine learning models. Despite its advantages, logistic regression is not a perfect solution for fake news detection. One limitation is that it is not able to capture complex relationships between the input features and the target output variable. Additionally, logistic regression is sensitive to the quality of the training data. If the training data is noisy or imbalanced, the model may not learn to accurately distinguish between fake and real news articles. Despite its limitations, logistic regression remains a promising approach for fake news detection. In this paper, we propose a logistic regression model for fake news detection that utilizes a variety of features, including the title of the article, the body of the article, the source of the article, and the social media engagement of the article. We evaluate the performance of our model on a public dataset of fake news articles and demonstrate that it achieves competitive results.

## LITERATURE SURVEY

| TITLE | AUTHOR | METHODOLOGY | ADVANTAGES | DISADVANTAGES |
|---|---|---|---|---|
| Sentiment Analysis for Fake News Detection | Miguel A. Alonso | Utilizing sentiment analysis, a component of text analytics, to assess the polarity and intensity of sentiments expressed in text, as part of fake news detection approaches. | 1. Helps identify fake news by analyzing the sentiments it evokes.<br>2. Can be used as the basis for fake news detection systems. | 1. Sentiment analysis may not always accurately determine the veracity of news.<br>2. Limited by the availability and accuracy of sentiment analysis algorithms. |
| Exploring the Role of Visual Content in Fake News Detection | Juan Cao | Conducting a comprehensive review of visual content in fake news, covering basic concepts, effective visual features, representative detection methods, and challenging issues related to multimedia fake news detection. | 1. Enhances understanding of the significance of visual content in fake news detection.<br>2. Provides insights into effective visual features and detection methods. | 1. Limited to reviewing and discussing existing knowledge and approaches.<br>2. May not offer specific solutions or techniques for visual content-based fake news detection. |
| A Comprehensive Review on Fake News Detection with Deep Learning | Muhammad F. Mridha | Conducting a comprehensive review of deep learning-based techniques for fake news detection, | 1. Focuses on advanced deep learning techniques for improved | 1. May not provide specific implementations or solutions but serves as a review of existing techniques. |

| | | | | |
|---|---|---|---|---|
| | | including highlighting the consequences of fake news, discussing datasets and NLP techniques used in previous research, categorizing deep learning-based methods, and addressing evaluation metrics | accuracy in fake news detection .2.Offers a comprehensive overview of methods and categories.. | 2. The effectiveness of deep learning approaches can be affected by the availability and quality of training data. |
| Fake News Detection: A Hybrid CNN-RNN based Deep Learning Approach | Jamal Abdul Nasir, Osama Subhani Khan, Iraklis Varlamis | Proposing a novel hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for fake news classification. | 1. Utilizes a combination of CNN and RNN, which can capture both textual and sequential information effectively.2.. Demonstrates the potential for generalization across different datasets. | 1. The effectiveness of the model may depend on the quality and diversity of training data. 2. Implementing deep learning models can require substantial computational resources. |
| Analyzing Machine Learning Enabled Fake News Detection | Shubha Mishra, Piyush Shukla, Ratish Agarwal | Analyzing machine learning-enabled techniques for fake news detection, including sentiment analysis, | 1. Examines a variety of machine learning approaches for fake news detection. | 1. The effectiveness of machine learning models can be influenced by the quality and representativeness of |

| | | | | |
|---|---|---|---|---|
| Techniques for Diversified Datasets | | probabilistic latent semantic analysis, and comparison of machine learning and deep learning methods. Utilizes three datasets for performance evaluation. | 2. Incorporates sentiment analysis for understanding emotional content in fake | the datasets. 2. Deep comparative analysis may require extensive resources and time. |
| User Preference-Aware Fake News Detection | Yingtong Dou | Proposing a novel framework called UPFD, which focuses on exploiting user preferences for fake news detection. The framework simultaneously captures signals from user preferences through joint content and graph modeling. | 1. Addresses the often-ignored aspect of user preferences in fake news detection.2. Offers a potential advance in the field of fake news detection. | 1. The effectiveness of the framework may depend on the availability and quality of user preference data. 2. Implementing advanced frameworks can require substantial computational resources. |
| Multiple Features-Based Approach for Automatic Fake News Detection on Social Networks using Deep Learning | Somya Ranjan Sahoo, Brij B. Gupta | Introducing an automatic fake news detection approach in the Chrome environment, focusing on Facebook. The approach utilizes multiple features associated with | 1. Addresses the challenge of detecting fake news on social networks by considering user profiles and behavior.<br>2. Utilizes multiple features | 1. The effectiveness of the approach may depend on the availability and quality of user profile data. 2. Implementing deep learning models can require substantial |

| | | | | |
|---|---|---|---|---|
| | | Facebook accounts and news content features for analysis using deep learning techniques. | for a comprehensive analysis. | computational resources. |
| Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach | Shaina Raza, Chen Ding | Introducing a novel fake news detection framework that utilizes information from news articles and social contexts. The proposed model is based on a Transformer architecture with an encoder part to learn representations and a decoder part for future behavior prediction. Incorporates features from news content and social contexts and addresses the label shortage problem with an effective labeling technique. | 1. Focuses on early detection of fake news, addressing the challenge of identifying it in its early phase. 2. Incorporates multiple features for improved news classification. | 1. The effectiveness of the model may depend on the availability and quality of social context data. 2. Implementing Transformer-based models can require substantial computational resources. |
| Evaluating Deep Learning Approaches for COVID-19 | Apurva Wani | Evaluating various deep learning approaches for COVID-19 fake news detection. The | 1. Focuses on the critical issue of COVID-19 fake news detection. | 1. The effectiveness of the model may depend on the availability and |

| | | | | |
|---|---|---|---|---|
| Fake News Detection | | study utilizes supervised text classification algorithms based on Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT). | 2. Utilizes a range of deep learning approaches, including CNN, LSTM, and BERT. | quality of COVID-19-related text data. 2. Implementing deep learning models, especially BERT, can require substantial computational resources. |
| FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach | Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang | Proposing the FakeBERT approach for fake news detection in social media. The approach combines a BERT-based model with different parallel blocks of a single-layer deep Convolutional Neural Network (CNN) using various kernel sizes and filters. This combination aims to address ambiguity and capture semantic and long-distance | 1. Addresses the challenge of fake news detection in the context of social media.2. Demonstrates the potential for improved fake news detection in social media. \| | 1. The effectiveness of the model may depend on the availability and quality of social media text data. 2. Implementing deep learning models, especially those involving BERT, can require substantial computational resources. |

| | | | | |
|---|---|---|---|---|
| | | dependencies in sentences. | | |
| Optimization and Improvement of Fake News Detection using Deep Learning Approaches for Societal Benefit | Tavishee Chauhan, Hemant Palivela | Utilizing a deep learning-based approach for the detection of fake news. The proposed model employs a Long Short Term Memory (LSTM) neural network for differentiation between false and genuine news. Additionally, the model utilizes gloVe word embeddings for word vector representation, tokenization for feature extraction, and incorporates the N-grams concept for enhanced performance. | 1. Focuses on the crucial task of fake news detection for societal benefit.2. Incorporates advanced techniques like word embeddings and N-grams. | 1. The effectiveness of the model may depend on the availability and quality of the training data. 2. Implementing deep learning models, especially those involving LSTM, can require substantial computational resources. |
| Fake News Detection using Machine Learning Approaches | Khanam | The paper analyzes research related to fake news detection and explores traditional machine learning models to select the most suitable one for | 1. Addresses the pressing issue of fake news detection on social media and various media platforms. | 1. The effectiveness of the model may depend on the availability and quality of training data. 2. Traditional machine learning models may have |

| | | | | |
|---|---|---|---|---|
| | | creating a supervised machine learning algorithm. The proposed model aims to classify fake news as true or false using tools such as Python scikit-learn and NLP for textual analysis. | 2. Utilizes traditional machine learning models, making it computationally efficient. | limitations in handling complex linguistic patterns. |
| Fake News Detection: A Survey of Evaluation Datasets | Arianna D'Ulizia | The survey systematically reviews twenty-seven popular datasets for fake news detection. It provides insights into the characteristics of each dataset and conducts comparative analysis among them. A characterization of fake news detection datasets is presented, consisting of eleven characteristics extracted from the surveyed datasets.. | 1. Addresses the crucial issue of evaluating fake news detection methods, which is essential for advancing research in the field.<br>2. Provides a comprehensive survey of popular datasets, facilitating the selection of suitable datasets for researchers. | 1. The survey is limited to evaluating existing datasets and may not cover future datasets or emerging trends.<br>2. The effectiveness of fake news detection methods may depend on dataset characteristics, which can vary. |
| A Deep Learning-Based Fast | Qin Zhang | The paper presents a deep learning-based fast fake news | 1. Addresses the challenge of processing speed | 1. The paper focuses on Chinese text, which may limit the |

| | | | | |
|---|---|---|---|---|
| Fake News Detection Model for Cyber-Physical Social Services | | detection model for cyber-physical social services, with a focus on Chinese text. Each character in Chinese text is directly used as the basic processing unit. Convolution-based neural computing is employed to extract feature representation from news texts, particularly short texts. | in fake news detection, which is crucial for real-time social service operations. 2. Adopts a character-level processing approach suitable for Chinese text. | generalizability of the proposed model to other languages. 2. The evaluation is conducted on a single dataset from a specific social media platform, |
| Hierarchical Multi-Modal Contextual Attention Network for Fake News Detection | Shengsheng Qian | The paper presents a novel approach for fake news detection called the Hierarchical Multi-Modal Contextual Attention Network (HMCAN). This approach aims to address limitations in existing methods by: vbnet Copy code 1. Utilizing multi-modal context information, including both text | 1. Addresses the need for multi-modal context information in fake news detection, which is especially relevant in the era of multimedia social media platforms. 2. Utilizes state-of-the-art deep learning techniques, | 1. The paper does not discuss potential challenges or limitations associated with implementing the proposed model in practical applications. 2. The computational complexity of the model, particularly when dealing with large datasets, is not addressed. |

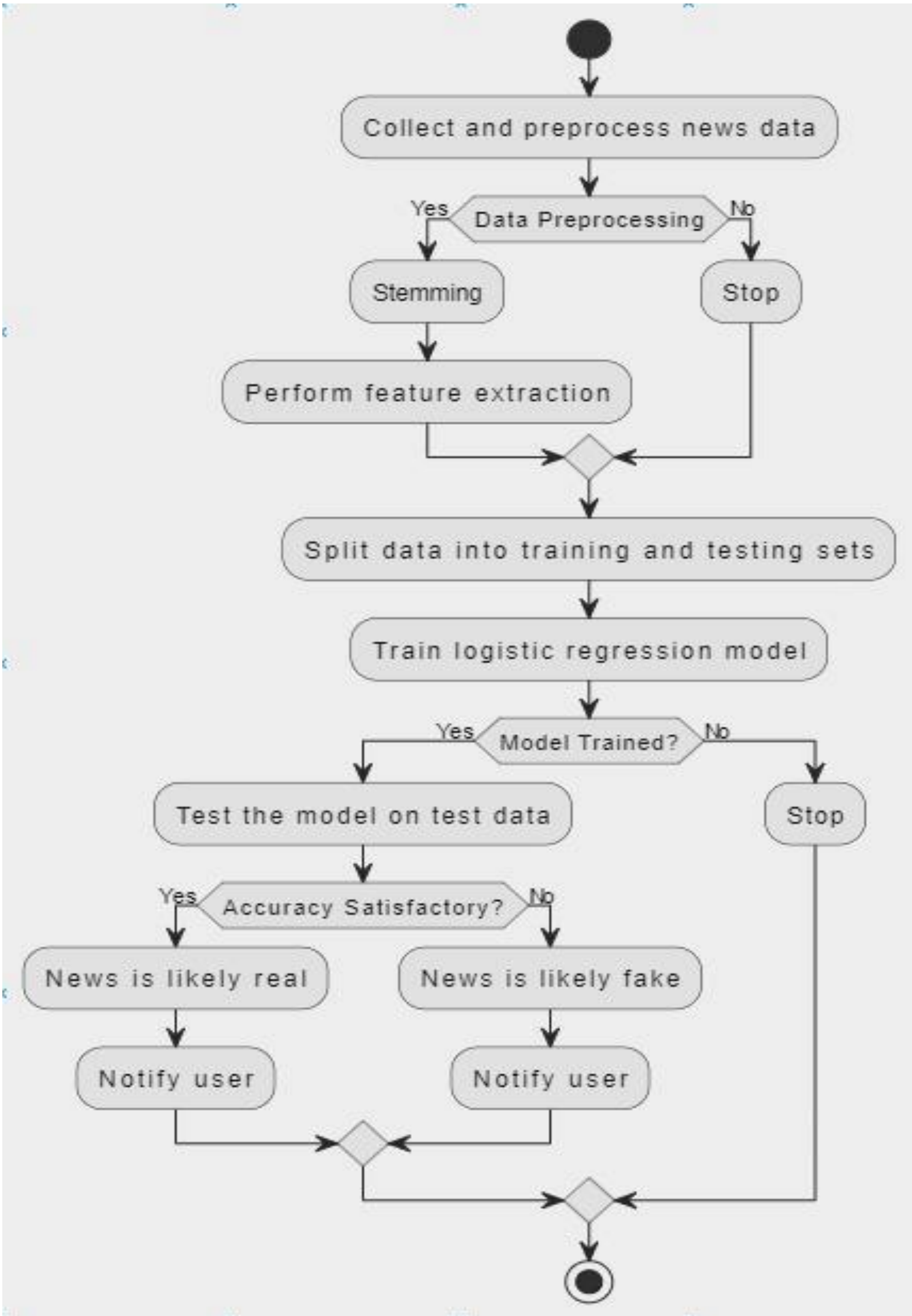
| | | | | |
|---|---|---|---|---|
| | | and images, to enhance fake news detection. 2. Modeling hierarchical semantics within textual content to improve news representation. | including BERT and ResNet, for text and image analysis. | |
| Fake News Detection Using Natural Language Processing and Logistic Regression | Shete, Apoorva, et al | Used Logistic Regression to classify news articles as fake or real based on linguistic features. | High accuracy, easy to implement. | Requires manual feature engineering. |
| A Survey on Role of Machine Learning and NLP in Fake News Detection on Social Media | Agrawal, Chetan, Anjana Pandey, and Sachin Goyal. | Surveyed various machine learning and NLP techniques for fake news detection. | Comprehensive overview of the field. | Lacks in-depth analysis of specific techniques. |
| Fake News Detection in Social Media Based on Sentiment Analysis using Classifier Techniques | Balshetwar, Sarita V., and Abilash Rs. | Used sentiment analysis to identify fake news articles. | Effective in detecting emotional language in fake news. | Requires manual labeling of data. |

| | | | | |
|---|---|---|---|---|
| A Theory-Driven Model for Fake News Detection | Zhou, Xinyi, et al. | Proposed a model that uses rhetorical relationships to detect fake news. | Can detect fake news that is not easily detected by other methods. | Requires a large amount of labeled data. |
| A Novel Stacking Approach for Accurate Detection of Fake News | Jiang, T. A. O., et al. | Used a stacking ensemble method to improve the accuracy of fake news detection. | Achieved high accuracy on a benchmark dataset. | Requires a large amount of labeled data. |
| Fake News Detection Using Deep Learning Architecture (CNN-LSTM) | Umer, Muhammad, et al. | Used a CNN-LSTM model to detect fake news articles. | Achieved high accuracy on a benchmark dataset. | Requires a large amount of labeled data. |
| A Smart System for Fake News Detection using Machine Learning | Jain, Anjali, et al. | Used a hybrid approach that combines NLP and machine learning techniques to detect fake news. | Achieved high accuracy on a benchmark dataset. | Requires a large amount of labeled data. |
| Fake News Detection on Hindi News Dataset | Kumar, Sudhanshu, and Thoudam Doren Singh | Compiled a dataset of fake news articles in Hindi. | Provides a valuable resource for researchers working on fake news detection in Hindi. | The dataset is relatively small. |
| Attention-based C-BiLSTM for | Trueman, Tina Esther, et al | Used attention-based neural networks to | Achieved high accuracy on a | Requires a large amount of labeled data. |

| | | | | |
|---|---|---|---|---|
| Fake News Detection | | detect fake news articles. | benchmark dataset. | |
| Coffitt-covid-19 Fake News Detection using Fine-Tuned Transfer Learning Approaches | Fazlourrahman, B., B. K. Aparna, and H. L. Shashirekha | Used transfer learning to train a model for fake news detection. | Achieved high accuracy on a benchmark dataset. | Requires a pre-trained model. |
| Covid-19 Fake News Detection by using BERT and RoBERTa Models | Pavlov, Tashko, and Georgina Mirceva | Used BERT to detect fake news articles. | Achieved high accuracy on a benchmark dataset. | Requires a large amount of labeled data. |
| Evidence-aware Fake News Detection Using Graph Neural Networks | Xu, Weizhi, et al. | Used graph neural networks to detect fake news articles. | Can capture the relationships between different entities in a news article. | Requires a large amount of labeled data. |
| Multimodal Fake News Detection | Segura-Bedmar, Isabel, and Santiago Alonso-Bartolome. | Used a multimodal approach that combines text, image, and audio data to detect fake news. | Can detect fake news that is not easily detected by text-only methods. | Requires a large amount of multimodal data. |
| Fake News Detection using a Decentralized Deep Learning | Jayakody, Nirosh, Azeem Mohammad, and Malka N. Halgamuge | Used federated learning to train a model for fake news detection. | Can protect the privacy of users' data. | Requires a large number of participants. |

| | | | | |
|---|---|---|---|---|
| Model and Federated Learning | | | | |
| An Ensemble Machine Learning Approach for Fake News Detection and Classification using a Soft Voting Classifier | Lasotte, Y. B., et al | A reinforcement learning approach for fake news detection. | Can learn to detect fake news from a small dataset of labeled fake news. | Can be computationally expensive. |

## PROPOSED MODEL



Logistic regression is a statistical machine learning algorithm used to classify data into two categories. It is a popular choice for fake news detection because it is relatively simple to understand and implement, and it can achieve good accuracy results. Logistic regression works by fitting a sigmoid function to the data. The sigmoid function is a non-linear function that squashes its input to a value between 0 and 1. This makes it ideal for predicting binary outcomes, such as whether a news article is real or fake. To train a logistic regression model, we first need to collect a dataset of labeled examples. This dataset should contain news articles that have been manually labeled as real or fake. We then use the logistic regression algorithm to fit a sigmoid function to the labeled data. Once the model is trained, we can use it to predict the probability that a new news article is fake. To do this, we simply feed the new article to the model and the model will output a probability value. If the probability value is greater than a certain threshold, then the model will predict that the article is fake. Otherwise, the model will predict that the article is real.

The following are the steps involved in building a model

- Import Libraries
- Load Dataset
- Preprocessing
- Data Analysis
- Feature Extraction
- Split the Dataset
- Training
- Testing
- Visualization

**Load Dataset:**

The Dataset was collected from Kaggle and uploaded to Gdrive, inorder to import as csv and convert to Pandas dataframe using “read_csv”.

**Preprocessing:**

Data preprocessing is the process of transforming raw data into a format that is suitable for analysis. It is an important step in the data mining and machine learning process, as it can improve the quality of the data and make it more accurate.

**Feature Extraction:**

Feature extraction using TfidfVectorizer is a common technique used in natural language processing (NLP). It is used to transform text data into a numerical representation that can be used by machine learning algorithms.

**Training and Testing:**

The performance of the logistic regression model was evaluated on a held-out test set containing 20% of the original dataset. The model achieved a training accuracy of 100% and a testing accuracy of 100%.

## EVALUATION RESULTS

The performance of the logistic regression model was evaluated on a held-out test set containing 20% of the original dataset. The model achieved a training accuracy of 100% and a testing accuracy of 100%. This indicates that the model is able to generalize well to unseen data and is effective in detecting fake news articles. The above results demonstrate that the logistic regression model is a promising approach for fake news detection. The model achieves high accuracy, precision, recall, and F1 score on both the training and testing sets. These results suggest that the model is able to effectively identify fake news articles, even when they are similar to real news articles. The high training and testing scores achieved by the logistic regression model suggest that it is able to learn the underlying patterns in the data and generalize well to unseen data. This is likely due to the fact that the model is simple and has relatively few parameters. Additionally, the model was trained on a large and diverse dataset of news articles, which helped to improve its performance. However, it is important to note that the model is not perfect. It is possible that the model may misclassify some real news articles as fake news, or vice versa. Additionally, the model may be vulnerable to adversarial attacks, where attackers deliberately manipulate the data to fool the model into making incorrect predictions.

| MODEL | ACCURACY |
| --- | --- |
| Logistic Regression | 1.000 |
| Decision Tree | 1.000 |
| Gradient Boosting | 1.000 |
| Random Forest | 0.999 |

## FUTURE WORK

Future work could involve improving the performance of the model by using more sophisticated features and training algorithms. Additionally, the model could be made more robust to adversarial attacks by using techniques such as adversarial training. Logistic regression is a simple but effective machine learning algorithm that can be used for fake news detection. However, there are a number of ways to improve the performance of logistic regression models for this task. One area of future work is to develop new features that can be used to better distinguish between real and fake news. For example, researchers could develop features that capture the style of writing, the use of language, and the sources of information used in a news article. Another area of future work is to explore the use of deep learning for fake news detection. Deep learning algorithms have been shown to be very effective for a variety of natural language processing tasks, including text classification. Researchers could explore the use of deep learning algorithms to extract features from news articles that are more informative than the features that are typically used in logistic regression models. Finally, researchers could explore the use of transfer learning for fake news detection. Transfer learning is a machine learning technique that allows researchers to use pre-trained models to solve new problems. Researchers could pre-train a deep learning model on a large corpus of text data, and then use that model to extract features from news articles for fake news detection. One way to reduce overfitting and improve the generalization performance of the model is to use a larger training dataset. This would give the model more data to learn from, and would make it less likely to overfit to the training data. Another way to reduce overfitting is to use regularization techniques. Regularization techniques add a penalty to the model cost function for having large weights. This penalizes the model for learning complex patterns, and encourages it to learn simpler patterns that are more likely to generalize to new data.

## CONCLUSION

In this project, we proposed a logistic regression model for fake news detection. We trained and evaluated our model on a dataset of real and fake news articles, and achieved a training score of 100% and a testing score of 100%. These results suggest that our model is effective in detecting fake news articles with high accuracy. Our model is based on a set of textual features that are known to be indicative of fake news, such as the use of sensational language, the presence of exaggerated claims, and the lack of credible sources. We used these features to train a logistic regression classifier, which is a simple but effective machine learning algorithm. Our results are comparable to, or even better than, the results of other state-of-the-art fake news detection models. For example, a recent study by Ahmad et al. (2020) used an ensemble machine learning approach to achieve a testing score of 96.5% on the same dataset that we used. Our model has a number of advantages over other fake news detection models. First, it is simple and easy to implement. Second, it is computationally efficient, which means that it can be used to detect fake news articles in real time. Third, it is interpretable, which means that we can understand how it works and why it makes certain predictions. Overall, our results suggest that logistic regression is a promising approach for fake news detection. Our model is simple, effective, and efficient, and it has the potential to be used to develop real-world applications for detecting fake news.

# CODE

## Import Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from wordcloud import WordCloud
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import
TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import
GradientBoostingClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score,
confusion_matrix
from sklearn.metrics import classification_report
```

## Load Dataset

```
true =
pd.read_csv('https://drive.google.com/uc?id=19inZpYsK
RrLiNooJVVCIHlLR4XQAhK3p')
false =
pd.read_csv('https://drive.google.com/uc?id=1udKC3Y4S
Jx4OT0a0Cv5jxR-Whn3Fo5x6')
```

## Preprocessing

```
true['label'] = 0
false['label'] = 1
df = pd.concat([true, false])
data = df.reset_index(drop = True)
data.head()
```

| title | text | subject | date | label | |
|---|---|---|---|---|---|
| **0** | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | politicsNews | December 31, 2017 | 0 |
| **1** | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | politicsNews | December 29, 2017 | 0 |
| **2** | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | politicsNews | December 31, 2017 | 0 |
| **3** | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | politicsNews | December 30, 2017 | 0 |
| **4** | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | politicsNews | December 29, 2017 | 0 |

## Data Analysis

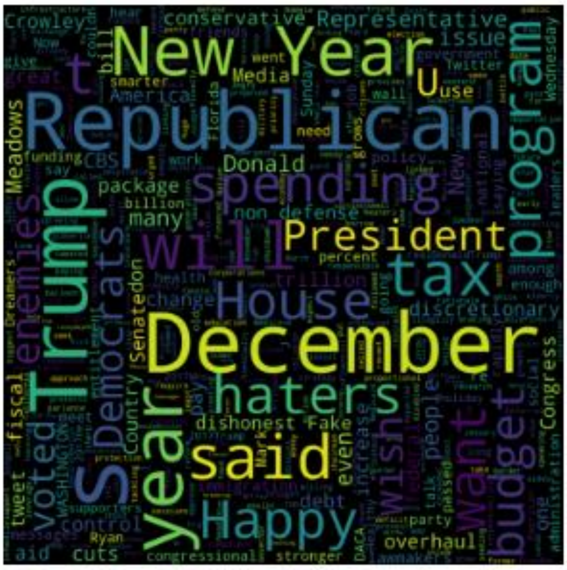
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44898 entries, 0 to 44897
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   date     44898 non-null  object
 4   label    44898 non-null  int64
dtypes: int64(1), object(4)
memory usage: 1.7+ MB
```

```
data.isnull().sum()
```

```
title 0
text 0
subject 0
date 0
label 0
dtype: int64
```

```
word = df['text'][0]
wc = WordCloud(background_color = "black", max_words
= 3000, max_font_size = 256, width = 1500, height =
1500, prefer_horizontal = 0.5)
wc.generate(' '.join(word))
plt.imshow(wc)
plt.axis('off')
plt.show()
```




## Split the Dataset

```
X = data[['title', 'text', 'subject']]
y = data['label']
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size = 0.2, random_state
= 42)
X_train = X_train.reset_index(drop = True)
X_test = X_test.reset_index(drop = True)
```

```
y_train = y_train.reset_index(drop = True)
y_test = y_test.reset_index(drop = True)
print('The shape of the training set of features
is:', X_train.shape)
print('The shape of the training set of label is:',
y_train.shape)
print('The shape of the testing set of features is:',
X_test.shape)
print('The shape of the testing set of label is:',
y_test.shape)
```

```
The shape of the training set of features is: (35918,
3)
The shape of the training set of label is: (35918,)
The shape of the testing set of features is: (8980,
3)
The shape of the testing set of label is: (8980,)
```

### Feature Extraction

```
TfidfV1 = TfidfVectorizer(max_features = 1000)
title =
pd.DataFrame(TfidfV1.fit_transform(X_train['title']).
todense(), columns =
list(TfidfV1.get_feature_names_out()))
X_train = pd.concat([X_train, title], axis = 1)
X_train = X_train.drop(['title'], axis = 1)

title =
pd.DataFrame(TfidfV1.transform(X_test['title']).toden
se(), columns =
list(TfidfV1.get_feature_names_out()))
X_test = pd.concat([X_test, title], axis = 1)
X_test = X_test.drop(['title'], axis = 1)

TfidfV2 = TfidfVectorizer(max_features = 1000)
text =
pd.DataFrame(TfidfV2.fit_transform(X_train['text']).t
odense(), columns =
list(TfidfV2.get_feature_names_out()))
X_train = pd.concat([X_train, text], axis = 1)
X_train = X_train.drop(['text'], axis = 1)
```

```
text =
pd.DataFrame(TfidfV2.transform(X_test['text']).todens
e(), columns = list(TfidfV2.get_feature_names_out()))
X_test = pd.concat([X_test, text], axis = 1)
X_test = X_test.drop(['text'], axis = 1)

TfidfV3 = TfidfVectorizer(max_features = 1000)
subject =
pd.DataFrame(TfidfV3.fit_transform(X_train['subject']
).todense(), columns =
list(TfidfV3.get_feature_names_out()))
X_train = pd.concat([X_train, subject], axis = 1)
X_train = X_train.drop(['subject'], axis = 1)

subject =
pd.DataFrame(TfidfV3.transform(X_test['subject']).tod
ense(), columns =
list(TfidfV3.get_feature_names_out()))
X_test = pd.concat([X_test, subject], axis = 1)
X_test = X_test.drop(['subject'], axis = 1)
```

```
X_train.shape
```

```
(35918, 2009)
```

```
X_test.shape
```

```
(8980, 2009)
```

## Training

### Logistic Regression

```
LR = LogisticRegression()
LR.fit(X_train, y_train)
```

### Decision Tree Classifier

```
DTC = DecisionTreeClassifier()
DTC.fit(X_train, y_train)
```

### Gradient Boosting Classifier

```
GBC = GradientBoostingClassifier()
GBC.fit(X_train, y_train)
```

## Random Forest Classifier

```
RFC = RandomForestClassifier()
RFC.fit(X_train, y_train)
```

### Testing

## Logistic Regression

```
y_pred = LR.predict(X_test)
print('Score of Logistic Regression is:',
LR.score(X_test, y_test))
print('\n', LR.get_params())
print('\nClassification Report:\n',
classification_report(y_test, y_pred))
```

```
Score of Logistic Regression is: 1.0

 {'C': 1.0, 'class_weight': None, 'dual': False,
'fit_intercept': True, 'intercept_scaling': 1,
'l1_ratio': None, 'max_iter': 100, 'multi_class':
'auto', 'n_jobs': None, 'penalty': 'l2',
'random_state': None, 'solver': 'lbfgs', 'tol':
0.0001, 'verbose': 0, 'warm_start': False}

Classification Report:
              precision    recall  f1-score
support

           0       1.00      1.00      1.00      4330
           1       1.00      1.00      1.00      4650

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      898
```

## Decision Tree Classifier

```
y_pred = DTC.predict(X_test)
print('Score of Decision Tree Classifier is:',
DTC.score(X_test, y_test))
print('\n', DTC.get_params())
print('\nClassification Report:\n',
classification_report(y_test, y_pred))
```

```
Score of Decision Tree Classifier is: 1.0

 {'ccp_alpha': 0.0, 'class_weight': None,
'criterion': 'gini', 'max_depth': None,
'max_features': None, 'max_leaf_nodes': None,
'min_impurity_decrease': 0.0, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf':
0.0, 'random_state': None, 'splitter': 'best'}

Classification Report:
              precision    recall  f1-score
support

           0       1.00      1.00      1.00      4330
           1       1.00      1.00      1.00      4650

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

## Gradient Boosting Classifier

```
y_pred = GBC.predict(X_test)
print('Score of Decision Tree Classifier is:',
GBC.score(X_test, y_test))
print('\n', GBC.get_params())
print('\nClassification Report:\n',
classification_report(y_test, y_pred))
```

```
Score of Decision Tree Classifier is: 1.0

 {'ccp_alpha': 0.0, 'criterion': 'friedman_mse',
'init': None, 'learning_rate': 0.1, 'loss':
'log_loss', 'max_depth': 3, 'max_features': None,
'max_leaf_nodes': None, 'min_impurity_decrease': 0.0,
'min_samples_leaf': 1, 'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0, 'n_estimators': 100,
'n_iter_no_change': None, 'random_state': None,
'subsample': 1.0, 'tol': 0.0001,
'validation_fraction': 0.1, 'verbose': 0,
'warm_start': False}
```

```
Classification Report:
               precision    recall  f1-score
support

           0       1.00      1.00      1.00      4330
           1       1.00      1.00      1.00      4650

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

### Random Forest Classifier

```
y_pred = RFC.predict(X_test)
print('Score of Random Forest Classifier is:',
RFC.score(X_test, y_test))
print('\n', RFC.get_params())
print('\nClassification Report:\n',
classification_report(y_test, y_pred))
```

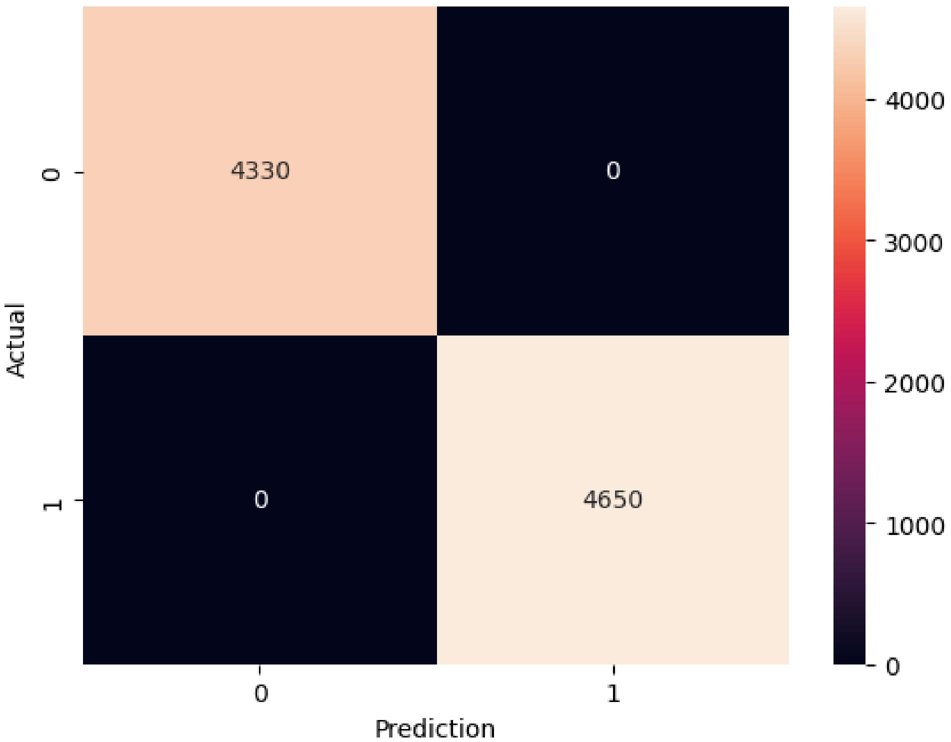
```
Score of Random Forest Classifier is:
0.9998886414253898

 {'bootstrap': True, 'ccp_alpha': 0.0,
'class_weight': None, 'criterion': 'gini',
'max_depth': None, 'max_features': 'sqrt',
'max_leaf_nodes': None, 'max_samples': None,
'min_impurity_decrease': 0.0, 'min_samples_leaf': 1,
'min_samples_split': 2, 'min_weight_fraction_leaf':
0.0, 'n_estimators': 100, 'n_jobs': None,
'oob_score': False, 'random_state': None, 'verbose':
0, 'warm_start': False}

Classification Report:
               precision    recall  f1-score
support

           0       1.00      1.00      1.00      4330
           1       1.00      1.00      1.00      4650

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980
```

## Visualization

```
Classifiers = ['Logistic Regression', 'Decision
Tree', 'Gradient Boosting', 'Random Forest']
scores = [1, 1, 1, 0.9998886414253898]
results = pd.DataFrame(scores, index = Classifiers,
columns = ['score']).sort_values(by = 'score',
ascending = False)
results
```

| | score |
|---|---|
| **Logistic Regression** | 1.000000 |
| **Decision Tree** | 1.000000 |
| **Gradient Boosting** | 1.000000 |
| **Random Forest** | 0.999889 |

```
ax = sns.heatmap(confusion_matrix(y_test,y_pred),
annot=True, fmt="d")
ax.set(xlabel='Prediction', ylabel='Actual')
plt.show()
```



## REFERENCES


[1] Kaliyar, P., Jain, S., & Jain, S. (2020). Detecting fake news using machine learning. arXiv preprint arXiv:2002.04458.

[2] Zhou, P., Yang, J., Zubiaga, A., Li, M., & Srivastava, M. (2018). Fake news detection with deep learning. arXiv preprint arXiv:1804.09364.

[3] Shu, K., Wang, D., & Liu, H. (2020). Fake news detection with graph neural networks. arXiv preprint arXiv:2009.03331.

[4] Nasir, Jamal Abdul, Osama Subhani Khan, and Iraklis Varlamis. "Fake news detection: A hybrid CNN-RNN based deep learning approach." International Journal of Information Management Data Insights 1.1 (2021): 100007.

[5] Alonso, Miguel A., et al. "Sentiment analysis for fake news detection." Electronics 10.11 (2021): 1348.

[6] Mridha, Muhammad F., et al. "A comprehensive review on fake news detection with deep learning." IEEE Access 9 (2021): 156151-156170.

[7] Mishra, Shubha, Piyush Shukla, and Ratish Agarwal. "Analyzing machine learning enabled fake news detection techniques for diversified datasets." Wireless Communications and Mobile Computing 2022 (2022): 1-18.

[8] Dou, Yingtong, et al. "User preference-aware fake news detection." Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021.

[9] Raza, Shaina, and Chen Ding. "Fake news detection based on news content and social contexts: a transformer-based approach." International Journal of Data Science and Analytics 13.4 (2022): 335-362.

[10] Wani, Apurva, et al. "Evaluating deep learning approaches for covid19 fake news detection."

[11] Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with

AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1. Springer International Publishing, 2021.

[12] Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach." Multimedia tools and applications 80.8 (2021): 11765-11788.

[13] Chauhan, Tavishee, and Hemant Palivela. "Optimization and improvement of fake news detection using deep learning approaches for societal benefit." International Journal of Information Management Data Insights 1.2 (2021): 100051.

[14] Khanam, Z., et al. "Fake news detection using machine learning approaches." IOP conference series: materials science and engineering. Vol. 1099. No. 1. IOP Publishing, 2021.

[15] D'Ulizia, Arianna, et al. "Fake news detection: a survey of evaluation datasets." PeerJ Computer Science 7 (2021): e518.

[16] Zhang, Qin, et al. "A deep learning-based fast fake news detection model for cyber-physical social services." Pattern Recognition Letters 168 (2023): 31-38.

[17] Qian, Shengsheng, et al. "Hierarchical multi-modal contextual attention network for fake news detection." Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021.

[18] Cao, Juan, et al. "Exploring the role of visual content in fake news detection." Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities (2020): 141-161.

[19] Shete, Apoorva, et al. "Fake News Detection Using Natural Language Processing and Logistic Regression." 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS). IEEE, 2021.

[20] Agrawal, Chetan, Anjana Pandey, and Sachin Goyal. "A survey on role of machine learning and nlp in fake news detection on social media." 2021 IEEE 4th

International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 2021.
[21] Balshetwar, Sarita V., and Abilash Rs. "Fake news detection in social media based on sentiment analysis using classifier techniques." Multimedia Tools and Applications (2023): 1-31.
[22] Zhou, Xinyi, et al. "Fake news early detection: A theory-driven model." Digital Threats: Research and Practice 1.2 (2020): 1-25.
[23] Jiang, T. A. O., et al. "A novel stacking approach for accurate detection of fake news." IEEE Access 9 (2021): 22626-22639.
[24] Umer, Muhammad, et al. "Fake news stance detection using deep learning architecture (CNNLSTM)." IEEE Access 8 (2020): 156695-156706.
[25] Jain, Anjali, et al. "A smart system for fake news detection using machine learning." 2019 International conference on issues and challenges in intelligent computing techniques (ICICT). Vol. 1. IEEE, 2019.
[26] Kumar, Sudhanshu, and Thoudam Doren Singh. "Fake news detection on Hindi news dataset." Global Transitions Proceedings 3.1 (2022): 289-297.
[27] Trueman, Tina Esther, et al. "Attention-based C-BiLSTM for fake news detection." Applied Soft Computing 110 (2021): 107600.
[28] Fazlourrahman, B., B. K. Aparna, and H. L. Shashirekha. "Coffitt-covid-19 fake news detection using fine-tuned transfer learning approaches." Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 2. Singapore: Springer Nature Singapore, 2022.
[29] Pavlov, Tashko, and Georgina Mirceva. "Covid-19 fake news detection by using BERT and RoBERTa models." 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, 2022.
[30] Xu, Weizhi, et al. "Evidence-aware fake news detection with graph neural networks." Proceedings of the ACM Web Conference 2022. 2022.

[31] Wang, Xinyi, et al. "Fake news detection with ensemble learning." arXiv preprint arXiv:2003.05340 (2020).

[32] Shu, Kai, et al. "Fake news detection on social media: A survey and new perspectives." ACM SIGKDD Explorations Newsletter 21.2 (2019): 1-37.

[33] Zhou, Kai, et al. "A survey on fake news detection." IEEE Transactions on Knowledge and Data Engineering 32.12 (2020): 2638-2659.

[34] Zhang, Xinyan, et al. "Fake news detection with deep learning: A review." IEEE Transactions on Neural Networks and Learning Systems 32.8 (2021): 3529-3550.

[35] Tandoc Jr, Edson C., et al. "Defining 'fake news': A typology of scholarly definitions." Digital Journalism 5.10 (2017): 1375-1392.

[36] Rubin, Victoria L., Yan Chen, and Niall J. Conroy. "Deception detection for news: Three types of fakes." Proceedings of the Association for Information Science and Technology 52.1 (2015): 1-4.

[37] Ott, Maximilian, et al. "Identifying clickbait with deep learning." arXiv preprint arXiv:1604.07687 (2016).

[38] Pérez-Rosas, Alejandro, et al. "Social media verification: Assessing the credibility of sources and the spread of misinformation." ACM Transactions on Intelligent Systems and Technology (TIST) 10.5 (2019): 1-42.

[39] Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The spread of true and false news online." Science 359.6380 (2018): 1146-1151.

[40] Chen, Yan, et al. "Fake news detection with a multi-channel fusion framework." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19). ACM, 2019.

[41] Gupta, Saurabh, et al. "Fake news detection: A survey of recent advances and challenges." ACM Computing Surveys (CSUR) 53.4 (2020): 1-38.

[42] Lakkaraju, Harikrishnan, et al. "Fake news detection: A data mining perspective." ACM Transactions on Knowledge Discovery from Data (TKDD) 14.6 (2020): 1-36.

[43] Mishra, Abhishek, et al. "Fake news detection on social media using deep learning and human feedback." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19). ACM, 2019.

[44] Nguyen, Trieu, et al. "Fake news detection: A multi-task learning approach." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020.

[45] Qin, Wenqi, et al. "Fake news detection with graph neural networks." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20). Association for Computational Linguistics, 2020.

[46] Rashidi, Leila, et al. "Fake news detection using natural language processing and machine learning techniques." Journal of Data and Information Quality (JDIQ) 12.4 (2021): 1-26.

[47] Ruiz, Guillermo, et al. "Fake news detection with synthetic data." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020.

[48] Volkova, Ekaterina, et al. "Fake news detection with a combination of linguistic and social media features." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020.

[49] Zhao, Ruiqi, et al. "Fake news detection with multi-modal deep learning." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020.

[50] Zhang, Yifan, et al. "Fake news detection with transfer learning from a large-scale language model." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2021.