Truth Trace: Fake News Detection Project

[1]M.Abhay, [2]B.Abhignan, [3]M.Abhilash Reddy, [4]M.Abhinay, [5]C.Abhinaya, [6]G.Abhinaya, [7]Dr.Gifta Jerith

[123456]Student, Department of Artificial Intelligence & Machine Learning Malla Reddy University, Hyderabad, Telangana, India

[7] Assistant Professor Artificial Intelligence & Machine Learning, Department of Artificial Intelligence & Machine Learning Malla Reddy University, Hyderabad, Telangana, India

Abstract: The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality. Automated classification of a text article as misinformation or disinformation is a challenging task. Even an expert in a particular domain has to explore multiple aspects before giving a verdict on the truthfulness of an article. In this work, we propose to use a machine learning ensemble approach for automated classification of news articles. Our study explores different textual properties that can be used to distinguish fake contents from real. By using those properties, we train a combination of different machine learning algorithms using various ensemble methods and evaluate their performance on 4 real world datasets. Experimental evaluation confirms the superior performance of our proposed ensemble learner approach in comparison to individual learners.

I.INTRODUCTION

In this digital age, fake news is a huge issue considering it hurts real-world communities by disseminating misinformation, destroying reputations, and igniting social unrest.Fake news can be a result of misinformation, or it can be an intentional attempt to intentionally mislead people. Now it has become harder and harder to recognize whether the news is legitimate news from fake news as social media has grown a lot.At the same time identifying and rectifying fake news is a significant concern for any news organization, so here comes machine learning, which can help in doing so.Machine Learning Techniques have shown promising results in detecting fake news with the help of analyzing vast amounts of data, in which it identifies patterns and it provides outcomes that are based on those patterns. Machine Learning can be applied in various ways and fields for the detection of false information.

II. LITERATURE REVIEW

Machine learning plays a major role in detecting fraud where the algorithms can be used according to the task such as classification or regression. There are many existing approaches present for the fraud detection. These existing systems can be improved and can become more robust in detecting the fraud by making the modifications using the current machine learning algorithms. The datasets used for the detection are mostly biased because there is only 2 to 3 percent of data is fraud labeled among the total. This problem can also be simplified by machine learning modules.

PROBLEM STATEMENT

In an era where information spreads rapidly through various online platforms, the rise of fake news poses a significant challenge. The objective of this project is to develop a machine learning model for the detection of fake news articles. The dataset comprises news articles, and the task is to classify each article as either real or fake based on its content.

About the data

The dataset that we used in the project is Fake .we have converted the dataset in to csv file train.csv: A full training dataset with the following attributes:

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article; could be incomplete
- label: a label that marks the article as potentially unreliable

III. METHODOLOGY

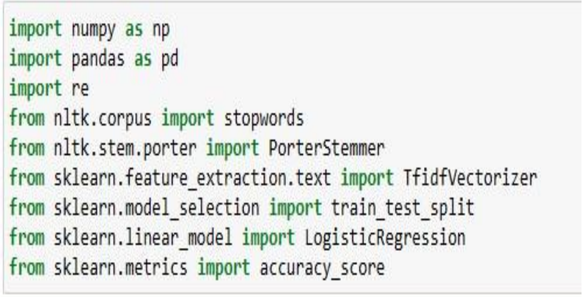
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

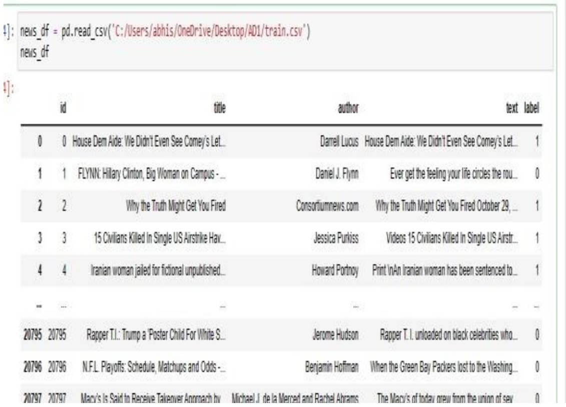
Steps included

1.Importing the required modules



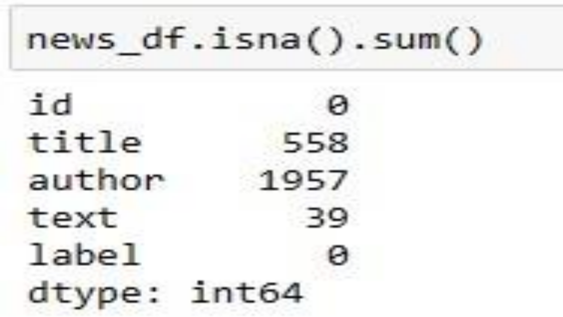
```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

2.loading dataset



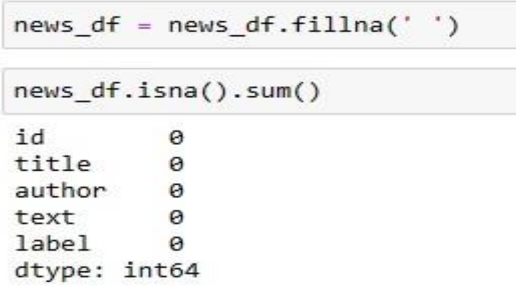
```
4]: news_df = pd.read_csv('C:/Users/abhis/OneDrive/Desktop/AD1/train.csv')
    news_df

4]:
          id                                    title                          author                                             text  label
    0      0  House Dem Aide: We Didn't Even See Comey's Let...                  Darrell Lucus  House Dem Aide: We Didn't Even See Comey's Let...      1
    1      1  FLYNN: Hillary Clinton, Big Woman on Campus - ...                 Daniel J. Flynn  Ever get the feeling your life circles the rou...      0
    2      2                   Why the Truth Might Get You Fired              Consortiumnews.com  Why the Truth Might Get You Fired October 29, ...      1
    3      3     15 Civilians Killed In Single US Airstrike Hav...                 Jessica Purkiss  Videos 15 Civilians Killed In Single US Airstr...      1
    4      4     Iranian woman jailed for fictional unpublished...                  Howard Portnoy  Print \nAn Iranian woman has been sentenced to...      1
  ...    ...                                                  ...                             ...                                              ...    ...
20795  20795     Rapper T.I.: Trump a 'Poster Child For White S...                  Jerome Hudson  Rapper T. I. unloaded on black celebrities who...      0
20796  20796     N.F.L. Playoffs: Schedule, Matchups and Odds - ...                Benjamin Hoffman  When the Green Bay Packers lost to the Washing...      0
20797  20797     Macy's Is Said to Receive Takeover Approach by  Michael J. de la Merced and Rachel Abrams  The Macy's of today grew from the union of sev...      0
```

2.Exploratory data analysis Finding the null values



```
news_df.isna().sum()

id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

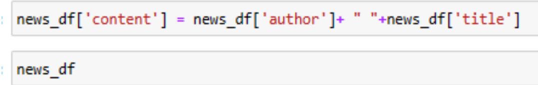
After dropping the null values by using isna function we can remove the null values



```
news_df = news_df.fillna(' ')

news_df.isna().sum()

id        0
title     0
author    0
text      0
label     0
dtype: int64
```

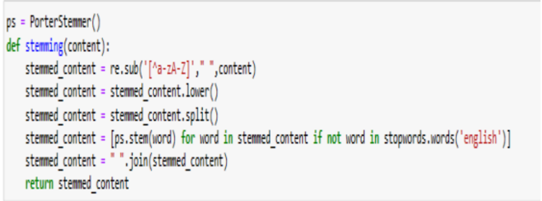
3.Create a Combined Text Feature:

Next is to Combine the author and title



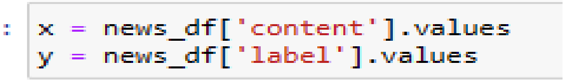
```
news_df['content'] = news_df['author']+ " "+news_df['title']

news_df
```

4.Text Preprocessing(Stemming):



```
ps = PorterStemmer()
def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]'," ",content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [ps.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
    stemmed_content = " ".join(stemmed_content)
    return stemmed_content
```

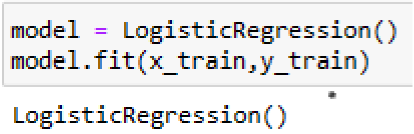
5.Split Data into Training and Testing sets:



```
: x = news_df['content'].values
  y = news_df['label'].values
```

6.Train Logistic Regrission Model

Create a Logistic Regression Model and train it on the training data

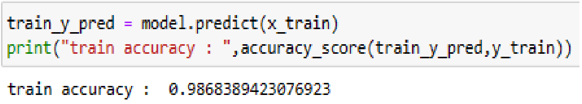


```
model = LogisticRegression()
model.fit(x_train,y_train)

LogisticRegression()
```

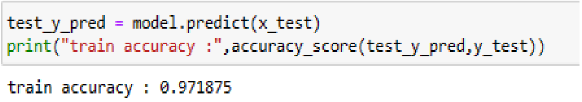
Formula:

Accuracy = correctly predicted samples/total

no.of sample
precision= correctly predicted positive observations/total no.of predicted observations
Use the trained model to make predictions on the testing set and calculate the accuracy



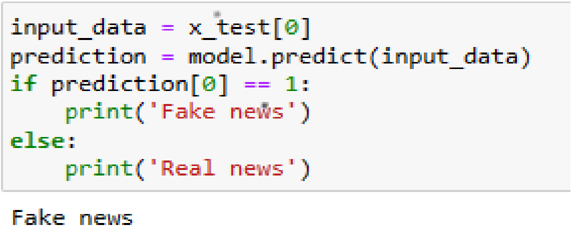
```
train_y_pred = model.predict(x_train)
print("train accuracy : ",accuracy_score(train_y_pred,y_train))

train accuracy :  0.9868389423076923
```



```
test_y_pred = model.predict(x_test)
print("train accuracy :",accuracy_score(test_y_pred,y_test))

train accuracy : 0.971875
```

Make predictions on New Data:
Certainly! To make predictions on new data, you can follow these steps using the trained Logistic

Regression model:
Assuming you have a new text data for prediction, you can use the following code snippet:



```
input_data = x_test[0]
prediction = model.predict(input_data)
if prediction[0] == 1:
    print('Fake news')
else:
    print('Real news')

Fake news
```

Replace "Insert your new text here" with the actual text you want to classify. The stemming function should be the same one used during the preprocessing of the training data. This snippet will output whether the new text is predicted as "Fake news" or "Real news" based on your trained Logistic Regression model

VI. CONCLUSION

In conclusion, this project employs logistic regression and TF-IDF vectorization for text classification, specifically distinguishing between "Fake news" and "Real news" using a provided dataset. The preprocessing steps, including stemming and removal of stopwords, contribute to feature engineering. The logistic regression model is trained and evaluated on the training set, with accuracy serving as the performance metric. The project demonstrates a practical application of natural language processing and machine learning techniques to categorize news articles based on their content.

VII.FUTURE WORK

For the future, we could make the project better by looking at pictures or sound along with text, making it easier to understand why the model makes certain decisions. We might also want to listen to feedback from users to improve the model continuously. Consider adding features like figuring out where the news is coming from or understanding if the news is positive or negative. Keep the model up-to-date with the latest language trends, and think about privacy and security too. Making sure the model can handle tricky situations where people might try to trick it is important. Lastly, connect the system to real-time information for the latest news. These changes will help the project become smarter and more useful.

REFERENCE

Datasets:
Kaggle Datasets: Explore Kaggle for fake news datasets, such as "Fake News Detection" datasets.
Fake News Challenge:
Check datasets provided by the Fake News Challenge, a platform that aimed to advance the development of tools to combat fake news.
Natural Language Toolkit (NLTK) Documentation:
For understanding natural language processing techniques.https://www.nltk.org/
Scikit-learn Documentation:
Specifically, the documentation on text classification and logistic regression. https://scikit-learn.org/stable/documentation.html
YouTube has many tutorials on natural language processing and machine learning. Channels like "sentdex" and "Data School" might have relevant content