Project Report

(PROJECT TERM JANUARY- MAY 2024)

Fake News Analysis Using Logistic Regression

SUBMITTED BY

Kamma Sai Pujitha (12209373)

Muhammed Suhail (12209097)

Srirag G(12208442)

SECTION K22RS

COURSE CODE INT254

UNDER THE GUIDENCE OF Prof. Rajan Kakkar

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

DECLERATION

I am here to declare that the project named “Fake News Analysis Using Logistic Regression is an authentic record of my own work carried out as requirements of the project for the award of B .Tech degree in the field of Computer Science and Engineering from Lovely Professional University, Phagwara , Punjab, under the supervision of Prof. Rajan Kakkar , during January to May 2024.All the information enhanced in this project report is based on my own aiming work and honest.

Kamma Sai Pujitha 30th March 2024
12209373

CERTIFICATE

This is to certify that the declaration statement made by this student is correct to the best of my knowledge and belief. He has completed this Project under my guidance and Supervision. The present work is the result of his original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfilment of the conditions for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara.

Prof. Rajan Kakkar

School of Computer Science and Engineering,

Lovely Professional University,

Phagwara, Punjab

Date: 30th March, 2024

Acknowledgement

It is with my immense gratitude that I acknowledge the support and help of my Professor, Prof. Rajan Kakkar, who has always encouraged me into this research. Without his constant guidance and help, this project would not have been a success for me. I am thankful to the Lovely Professional University, Punjab and the department of Computer Science without which this project would have not been an achievement. I also thank my family and friends, for their endless love and support throughout my life.

Table of Contents

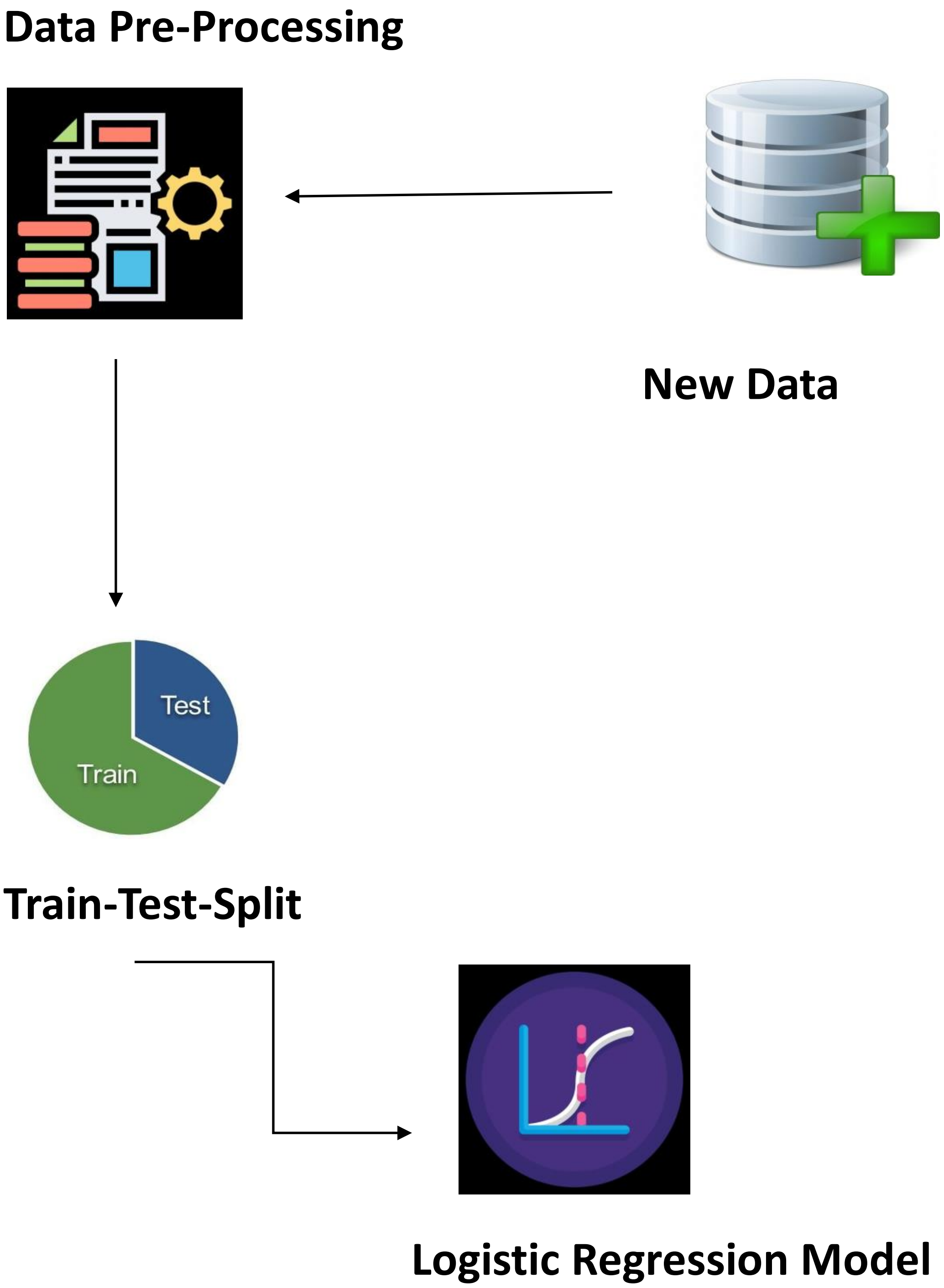
Abstract

Now a days due to the development of social networks , fake news become a major threat to the public .so it is an important goal to detect the fake news from real news . In this report , we present the use of natural language processing techniques to identify ‘false news’ , which were misleading the public . This method uses Logistic Regression classification model to predict wheather the news is labeled as real or fake .And the received results suggest , that fake news Analysis problem can be addressed with machine learning methods.

Keywords:

Machine Learning , Logistic Regression , Accuracy

Work Plan



Data Pre-Processing

New Data

Train-Test-Split

Logistic Regression Model

Introduction

Machine Learning is a field of study in artificial intelligence . It approaches have been applied to many branches including Email Filtering , Agriculture , Speech Recognition , Computer Vision , Natural Language Processing and medicine.

7.1)Natural Language Processing (NLP)

Natural language processing is the ability of a computer to understand human language . It is a component of artificial Intelligence that centre on instructing computers to efficiently examine massive volumes of natural language data. In the fields of linguistics, information engineering, computer science, and artificial intelligence, natural language processing (NLP) studies how computers interact with human languages. Its major aim is to instruct computer programmers in how to study and examine vast amounts of natural language.

7.2)Fake News Analysis

Along with the rising number of population useage of social media platforms have been increased, And false news has become a severe problem in recent days.

Knowing whether the news is fake or real .The field of fake news Analysis has rapidly progressed as a result of researchers and engineers developing a number of techniques to identify and takes up arms against misleading information.

7.3)Objective

Our project's primary aim is to determine the accuracy of news in order to determine if it is real or fake. the development of a machine learning model that would allow us to recognise false information. It can be difficult and difficult to identify fake news only based on its content since it is intentionally produced to influence readers to believe false information. By applying a range of methods and models, machine learning makes it easy to detect false news. Additionally, to examine the relationship between two words, we will apply deep learning-based NLP. We will also eliminate stop words using this method.

1.4)Abbreviations

ML = Machine Learning

LR = Logistic Regression

NLP = Natural Language Processing

NLTK = Natural Language Tool Kit

RE = Regular Expression

DATASET DESCRIPTION

Dataset contain information related to the news headlines and articles ,along with some subheadings like id, title, author, text, label. And binary label indicating whether the considered article is fake or not (0 for real news, 1 for fake news).Each and every row that is present in the dataset is a separate article or headline entry.

1.id : unique id for a news article or headline

2.title : the title of a new article or headline

3.author : author of the news article or headline

4.text : the text of the article(whether it would be complete or incomplete text)

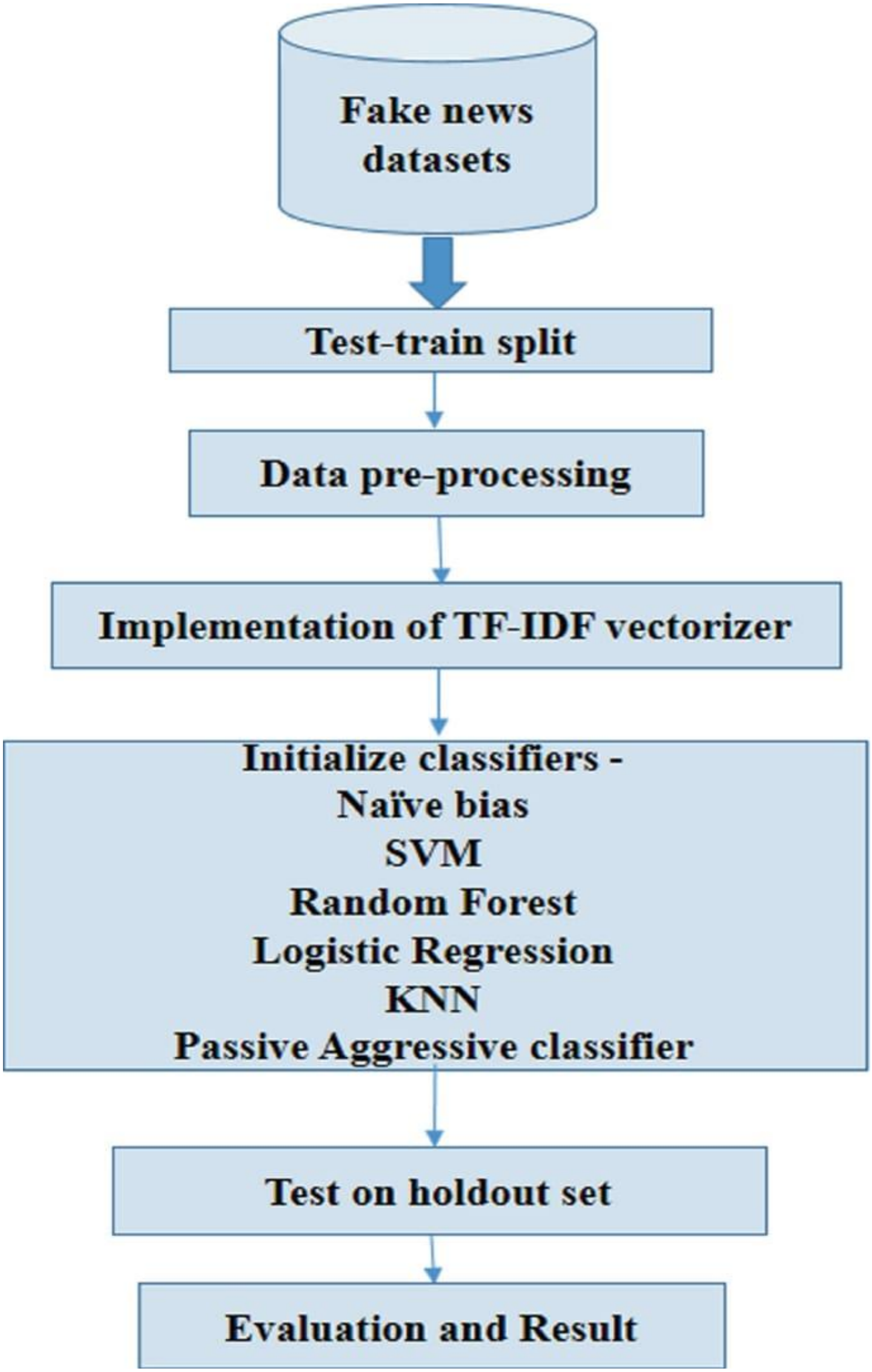
5.label : a label will mark whether the news article is real or fake

1 : Fake news

0 : Real news

The whole dataset contains 20800 number of rows and 5 number of columns.

Example Flow Chart:



MODULES

1. **Data use**
2. **Data Preprocessing**
3. **Training**
4. **Accuracy**

MODULES DESCRIPTION

1. DATA USE

In this project we are using different packages and libraries to load and read the dataset we are using pandas library. By using pandas, we can read the .csv file and then we can display the shape of the dataset. We will be training and testing the data, when we use supervised learning it means we are labeling the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be preprocessing that is null values which are not readable are required to be removed from the dataset.

2. DATA PRE-PROCESSING

- Load the dataset of news items with their labels, whether they are true or false
- Clean the text by eliminating punctuation and stopwords
- Divide the dataset into training and testing sets.

Count Vectorization

- Count Vectorizer from the sklearn toolkit which will be used to transform the text data into numerical data.
- Fit the Count Vectorizer using the training set, then convert the data.
- Utilise the testing set to change the data.

3. Training the Models:

- Utilise the data that has been modified by Count Vectorizer to train a variety of models, including Naive Bayes, Logistic Regression etc.
- Fit the models using the training set.

- Use the testing set to predict the news article labels.

  Confusion Matrix

  - The confusion matrix displays the amount of true positives, true negatives, false positives, and false negatives for each model, allowing you to assess each one's performance.
  - Measurements like accuracy, recall, and F1-score may be calculated using the confusion matrix.

Step 6: Accuracy

- Determine each model's accuracy by comparing its predicted labels to its actual labels.
- The accuracy measures the proportion of news stories that were accurately identified as being true or false.
- Evaluate the accuracy of various models to find which one is most effective at spotting fake news.

RESULT AND EXPERIMENTAL ANALYSIS

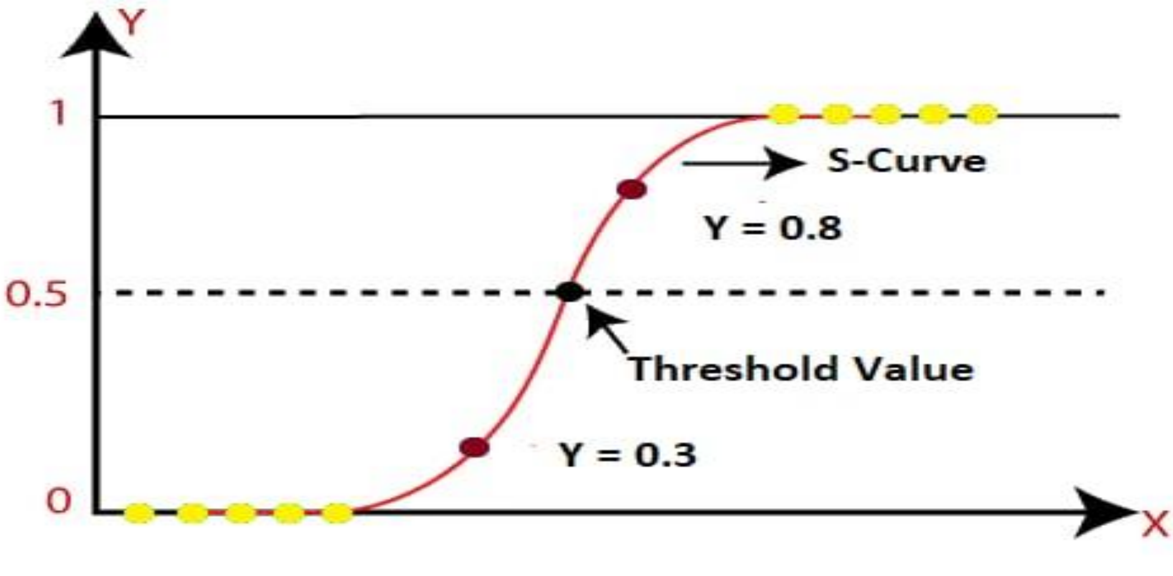
The model that we applied on our data

Logistic Regression:

In binary classification problems, the main aim is to frequently used method or model. With the use of a predict one of two outcomes, logistic regression is a sigmoid function, it converts the output of the linear regression into a probability value between 0 and 1, which can then be used to decide whether to classify data by applying a threshold.

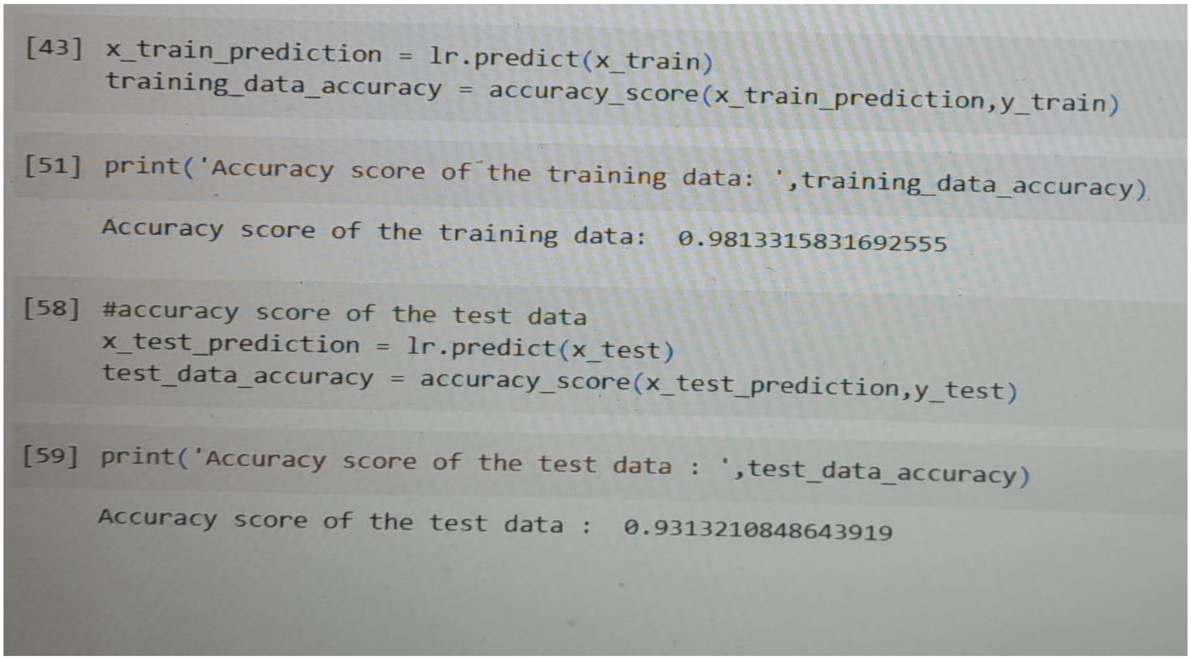
With applications in many areas, including spam filtering, credit scoring, and medical diagnosis, and this simple reliable algorithm may be taught well on the big datasets.

However,it depends on certain presumptions, such as the independence of the characteristics and linearity, it could not work well with highly coupled or nonlinear data.



Logistic regression

These are the Results from applying the Logistic Regression model:



```
[43] x_train_prediction = lr.predict(x_train)
     training_data_accuracy = accuracy_score(x_train_prediction,y_train)

[51] print('Accuracy score of the training data: ',training_data_accuracy)

     Accuracy score of the training data:  0.9813315831692555

[58] #accuracy score of the test data
     x_test_prediction = lr.predict(x_test)
     test_data_accuracy = accuracy_score(x_test_prediction,y_test)

[59] print('Accuracy score of the test data : ',test_data_accuracy)

     Accuracy score of the test data :  0.9313210848643919
```

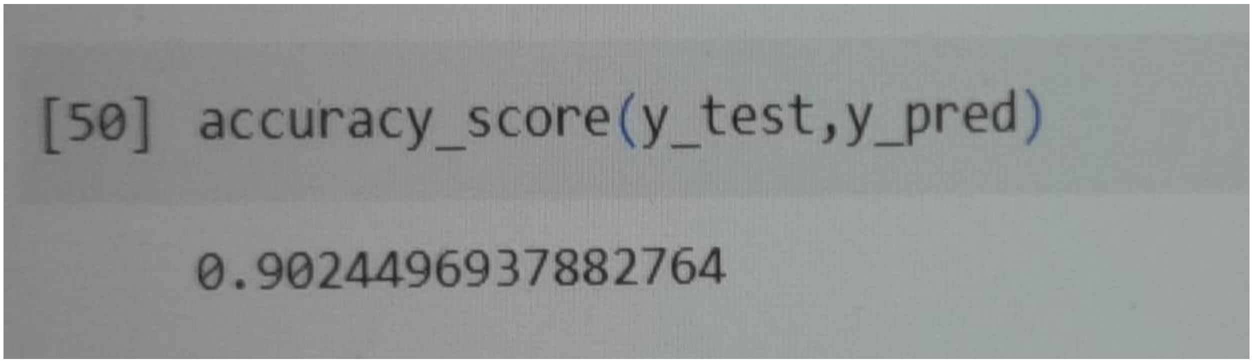
Multinomial Naive Bayes:

It is a very efficient and popular machine learning algorithm which is based on Bayes theorem.it is commonly used for text classification problems where we need to deal with discrete data.

Naïve Bayes is a probabilistic algorithm based on Bayes theorem .and the presence of one feature does not affect the presence of another.

Multinomial Naive Bayes is a probabilistic classifier to calculate probability distributions of the text data,it is well suited for data with features that represent frequencies of events in various natural language processing tasks.

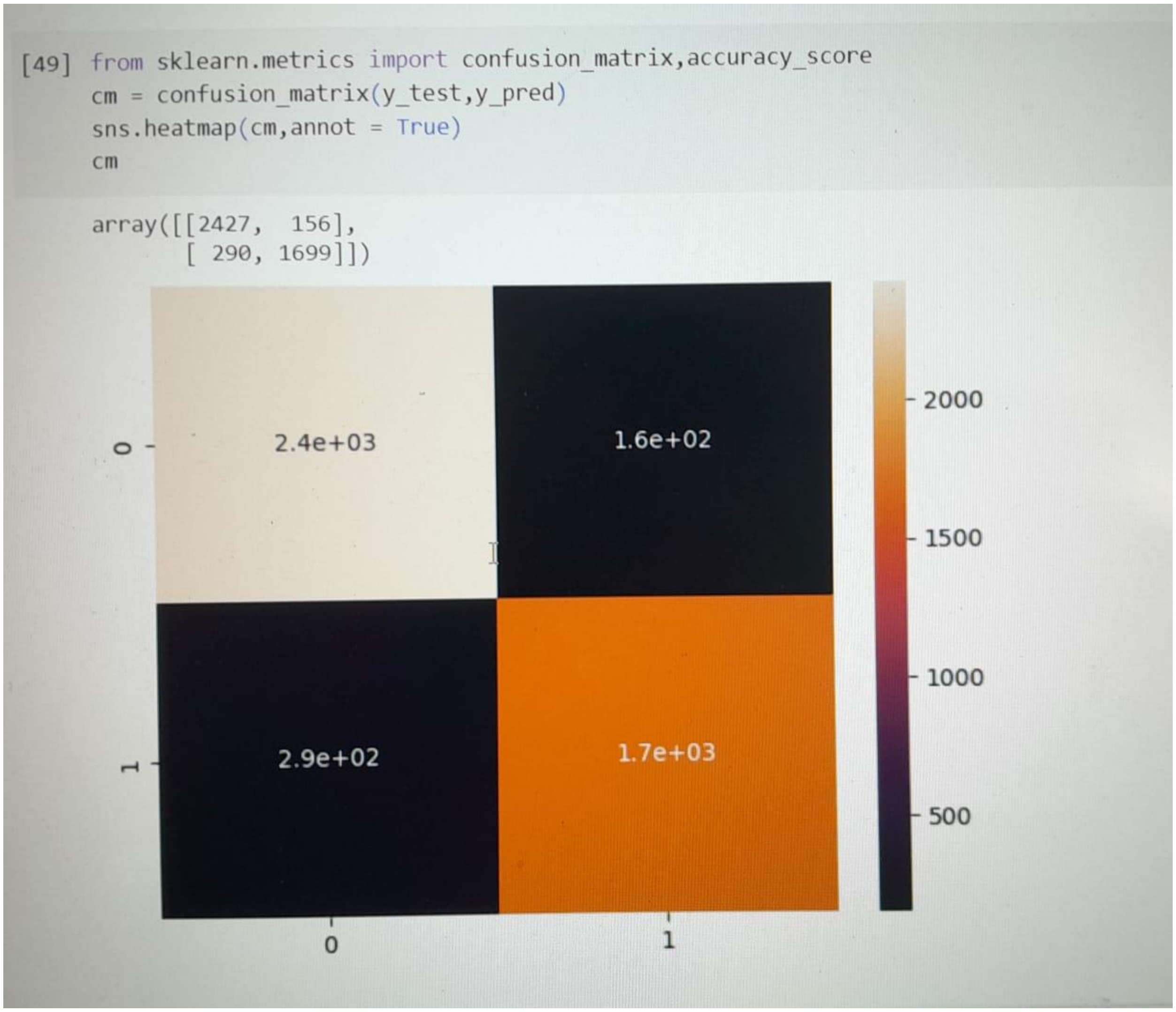
These are the Results from applying the Multinomial Naïve Bayes model:



```
[50] accuracy_score(y_test,y_pred)

     0.9024496937882764
```

Confusion Matrix:



```
[49] from sklearn.metrics import confusion_matrix,accuracy_score
     cm = confusion_matrix(y_test,y_pred)
     sns.heatmap(cm,annot = True)
     cm
```

```
array([[2427,  156],
       [ 290, 1699]])
```

Conclusion

Considering the accuracy scores , we were able to establish for the various models, we get to know that all of the models are doing a good job of identifying false news items. The Logistic Regression model notably achieved a very good accuracy of 93.13%, although the Multinomial Naïve Bayes model performed just slightly lower, at 90.24%.When we compare the both models logistic regression model was performed well than the multi naïve bayes model.

So, by Investigating different feature extraction and selection methods, classifier types, and ensemble approaches may also be useful to see whether even better results may be produced than the models that we have been worked on.

Accuracy of Logistic Regression model:

93.13%

Accuracy of Multinomial Naïve Bayes model:

90.24%

FUTURE SCOPE

Future research and advancement in the field of false news Analysis are definitely possible. Future efforts to identify false news may go in the following directions:

Including more subtle and varied aspects:

The most part, current ways for Analyzing false news depends on simple text-based traits like TF-IDF vectors or bag-of-words. Research in the future could concentrate on more complex and diverse aspects, such as sentiment analysis, multimedia analysis or network analysis.

Creating more interpretable models:

Already existing ways for detecting or analyzing fake news sometimes depends on complex machine learning algorithms that might be difficult to comprehend. In the future, it would be more beneficial to develop more intelligible models that might provide more information on how people make decisions.

Combining information from other resources:

In addition to social media, videos, and news articles, fake news is regularly spread through other media channels and platforms. The development of methods that can incorporate data from several sources may be crucial in the future to improve false news identification.

Adapting to shifting strategies:

It will be crucial for fake news Analyzing technologies to develop alongside the tactics used by those who create and spread it. For this, the detection methods might need to be regularly reviewed and improved

References

[1]. Haiden, L., & Althuis, J. (2018). The Definitional Challenges of Fake News

[2]. Helmstetter, S., & Paulheim, H. (2018, August). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274-277). IEEE.
[4]. Stahl, K. (2018). Fake News Detection in Social Media.

[3]. Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May). Automatic Online Fake News Detection Combining Content and Social Signals. In 2018 22nd Conference of Open Innovations Association (FRUCT) (pp. 272-279). IEEE

[4]. Parikh, S. B., & Atrey, P. K. (2018, April). Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.

[5]. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November). Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information

Science with Impact: Research in and for the Community (p. 82). American Society for Information Science

[6] A. S. A. Ahmed, A. Abidin, M. A. Maarof, and R. A. Rashid, "Fake news detection: A survey," IEEE Access, vol. 9, pp. 113051-113071, 2021. doi: 10.1109/ACCESS.2021.3104178

[7] S. Asghar, S. Mahmood, and H. Kamran, "Fake news detection using machine learning: A survey," IEEE Access, vol. 9, pp. 57613-57639, 2021. doi: 10.1109/ACCESS.2021.3075392

[8] J. H. Kim, S. H. Lee, and H. J. Kim, "Fake news detection using ensemble learning with context and attention mechanism," IEEE Access, vol. 9, pp. 27569- 27579, 2021. doi: 10.1109/ACCESS.2021.3057736 [9] M. F. Hossain, M. M. Islam, M. A. H. Khan, and J. J. Jung, "Fake news detection using hybrid machine learning algorithms," IEEE Access, vol. 8, pp. 233350-233364, 2020. doi: 10.1109/ACCESS.2020.3041149 [10] S. S. Ghosh, A. Mukherjee, and N. Ganguly, "A multi-perspective approach to fake news detection," IEEE Intelligent Systems, vol. 35, no. 5, pp. 31-39, 2020. doi: 10.1109/MIS.2020.3012915