Fake News Detector with Real Time Web Scraping

Pritam Roy[1], Smarto Chowdhury[2], Saptarshi Sarkar[3], Swarnadeep Sen[4]

Department Of Electronics and Communication Engineering, Narula Institute of Technology, Agarpara

Abstract: This project, Fake News Detector, articulates the power of the Python programming language the detector is designed in a way so that it can differentiate the given texts into real or fake. Fake news affects societies across the globe this survey explores important work in this area, shedding light on the methods researchers have developed, what they do well, and where they fall short.

I. INTRODUCTION

In today's era where in just one click of the mouse button Information is accessible within the clutches of our hand, fake news has become a major challenge in day to day lives.

Misinformation in the form of fake news influences opinion of the public , can manipulate perceptions and may lead to social unrest. Thus with passing of days, the need for automated tools to distinguish and combat fake news has become a integral part in our life.

As with the advancement of technology, many difficulties are overcome to lead a better life. And with technical languages and skills which are a boon to the world, engineers are able to solve problems that rise in the real life.

This project, Fake News Detector, articulates the power of the Python programming language The detector is designed in a way so that it can to differentiate the given texts into real or fake .

By implementing a this model, this project aims in providing a trustworthy, robust , consistent and modular solution to help in the prediction of fake news. This journal tells the story of the project's journey—from research and development to its implementation. It highlights the methods we used, the challenges we faced along the way, and the results we were able to achieve.

II. METHODOLOGY

As we know, our Project is about a Fake News Detection System which combines a pre-trained "Machine Learning" model with "Real-Time Web Scraping" to determine news articles as "Real" or "Fake". Let us give a detailed breakdown about this exciting project.

A. Modules Used in the Project and their roles

1) *Joblib:* This module is used for loading a pre-trained machine learning model(fake_news_model.pkl) and the vectorizer(tfidf_vectorizer.pkl). These files store the trained machine learning pipeline which enables the system to make quick and efficient predictions

2) *Gradio :* Gradio is a very useful module which helps us by providing a user-friendly web interface. It simplifies the making of a Front-end for the tool which helps users interact even if they have much programming skills.

3) *Re:* re is also a very popular and useful module in Python it stands for "Regular Expressions". The main purpose of this module is to clean the texts by removing special characters, numbers and symbols such as #,$%. The reason for cleaning the texts is mainly to improve the accuracy and efficiency of machine learning models.

4) *Requests:* requests module fetches the HTML content of the URL provided by the user. It extracts the text content from webpages.

5) *BeautifulSoup(from bs4) :* This module parses the HTML content and helps in extracting meaningful data like paragraphs(<p> tags). This process makes it easier to extract text from a Web Page's HTML structure.

6) *NLTK:* NLTK is also a popular module in Python which is used for reading text segments and preparing it for analysis by removing common and irrelevant words(stop-words) and reducing them to their base form(stemming).

7) *Matplotlib.pyplot:* Matplotlib module in python helps us in Data Visualization by allowing us to create exciting graphs such as pie charts and helps us explain the decision with a visual breakdown.

B. Step-by-Step Working of the Code

1) *Loading the Model:* Joblib module loads a Logistic Regression Model(fake_news_model.pkl) and a TfidfVectorizer and these tools helps us to transform text data into numerical data and classify it.
2) *Fetching Content from a URL:* The Webpage is downloaded with the help of fetch_news_content function in requests and BeautifulSoup extract the textual contents from <p> tags.
3) *Pre-processing the Content:* The re module cleans the text and the stop-words in them are removed with the help of nltk's PorterStemmer.
4) *Making Predictions:* TfidfVectorizer converts the cleaned text into a numerical format and passes it to the Logistic Regression model for classification. The model predicts whether the news is "Fake" or "Real" and provides probabilities for each class.
5) *Generating Output:* The System returns a result in text format "Real" or "Fake" and also shows us a Pie-chart for showing the accuracy levels of the output.
6) *Interactive User Interface :* We have used Gradio module for making an interactive user interface which comes with a text box where we should paste the URL of the news article and with a submit button for submitting the input.

C. About the Machine Learning Model

Logistic Regression: It is a Linear model which gives us binary classification(Real vs Fake). It is very simple, efficient and effective in case of text classification tasks and gives us output probabilities for each class.This model was trained on a dataset of labelled news articles. The TfidfVectorizer converted text into numerical features, capturing importance and frequency of every words. Logistic Regression learned the patterns associated with "Fake" and "Real" news**.**

III.   RESULT

Ten percent of the data was allocated for testing, while the remaining 80% was utilized for training the models. Logistic Regression was employed for classification, and their performance is given below figures our employed logical regression model achieved   50% accuracy rate here in the following figure we have

```
Accuracy: 50.00%

Classification Report:
                precision    recall  f1-score   support

  Real News       0.50      1.00      0.67         1
  Fake News       0.00      0.00      0.00         1

   accuracy                           0.50         2
  macro avg       0.25      0.50      0.33         2
weighted avg       0.25      0.50      0.33         2
```

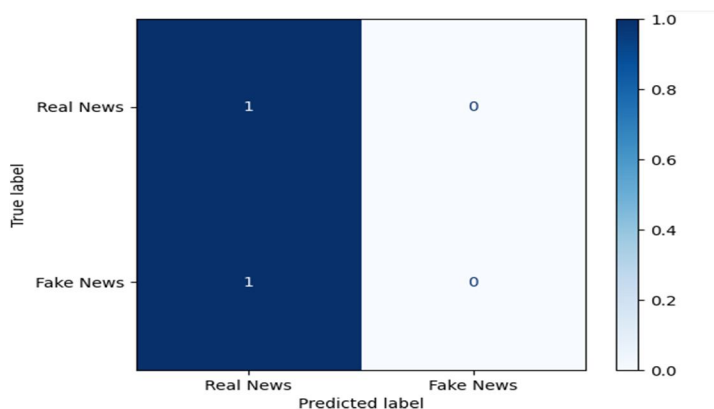Fig 1: Logical Regression Model Summary

Fig 2: Bar Diagram For Logical Regression Model

Now we are going to test our webpage using some urls available in the internet to see whether the news is real or not, sample url is given in reference section
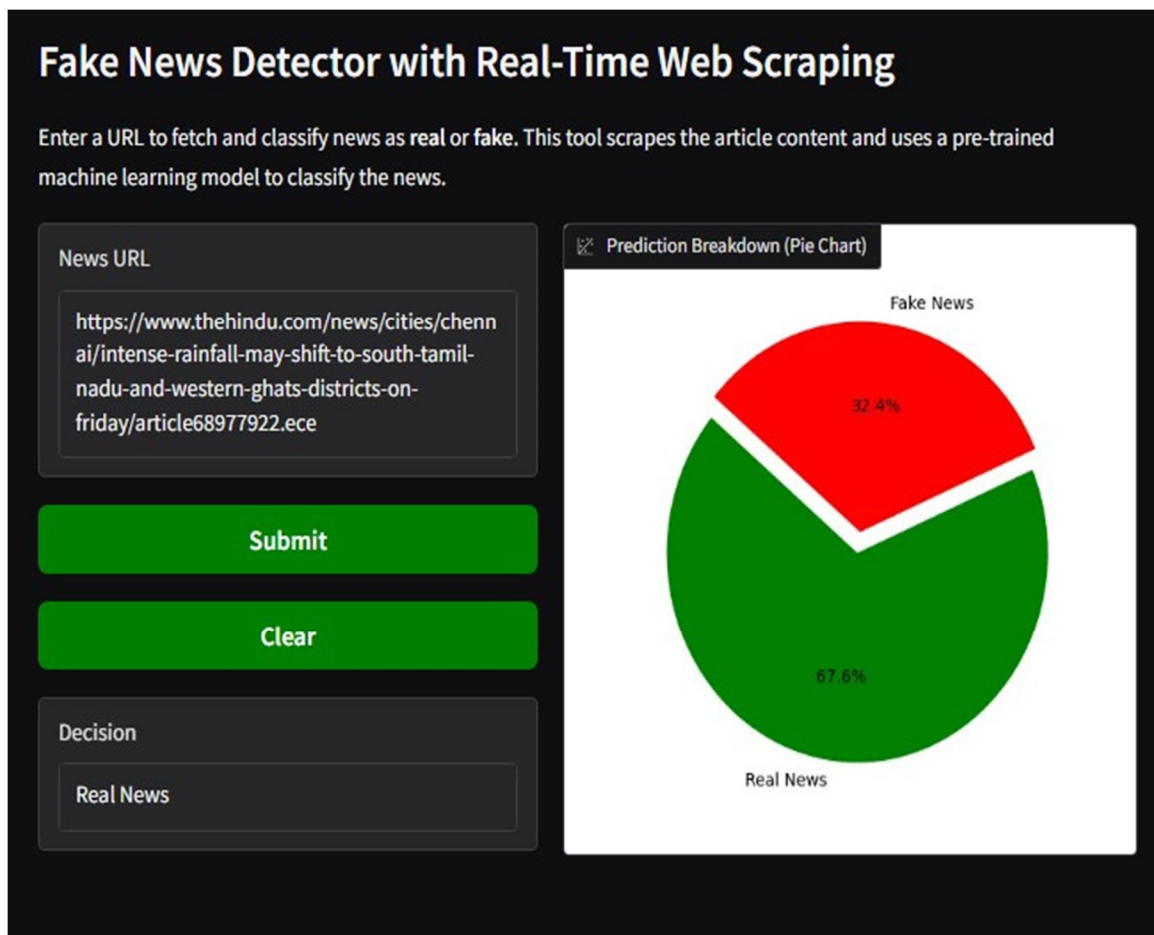
A. Conclusion

Fig 3: Visual Representation Of Working Of Web Page

IV. CONCLUSIONS

The spread of false news is a serious threat to information integrity in the current digital era. The techniques used to disseminate false information also evolve with technology. As a result, detecting false news effectively becomes a critical topic of attention. By employing modern algorithms and artificial intelligence, we can identify and mitigate/false information*more effectively. But technology isn't the answer on its own. Media literacy and critical thinking are essential for enabling people to distinguish between false and reliable information. These abilities can be developed through educational programs and public awareness initiatives, assisting individuals in challenging and validating the information they come across. We can create more effective tools and tactics to stop the spread of false information and protect the accuracy of the data in our

A. Future Scopes

Some of the challenges arising from Understanding fake news are as follows; Challenging: Dynamicity, changing strategies of fake news, Lack of diverse data. Looking at the future work, the main areas of improvement include improving model stability, using a combination of data types (image, video) and building real-time detection mechanisms. They are currently creating sophisticated models that can search and analyze decimal patterns of fake news to detect them in real-time, offering consumers push notifications when they are watching content on the internet. Integrated platforms' goal is to establish a wide interconnected network for fake news detection on different platforms and social media. There are also measures to promote user education through tools that can help the person himself determine the reliability of the information.

In addition, collaboration with independent fact-checking agencies and other news organizations is being done to improve credibility and reliability of the assessments. Both of these advancements are designed to enhance disinformation detection and filtering efficacy.

V. ACKNOWLEDGMENT

REFERENCES

[1] https://dl.acm.org/doi/10.1145/3313991.3314008
[2] https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211
[3] https://ieeexplore.ieee.org/abstract/document/9620068
[4] https://www.javatpoint.com/fake-news-detector-using-python
[5] https://www.thehindu.com/news/cities/chennai/intense-rainfall-may-shift-to-south-tamil-nadu-and-western-ghats-districts-on-friday/article68977922.ece

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 ⊙ (24*7 Support on Whatsapp)