School of Computer Science

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
DEHRADUN, UTTARAKHAND

Project Report

on

Fake News Classification

Submitted by:

Sanya Gilhotra
R2142210702
500091032
B.Tech(H)CSE AI ML(3rd Year)

Under the guidance of

Prof. Touseef Iqbal

Assistant Professor

AI-ML Cluster, School of Computer Science (SoCS), UPES Bidholi

Introduction

In the digital age, fake news—deliberately misleading or false information presented as news—has become widespread . This phenomenon undermines trust in traditional media sources, causing scepticism and confusion among the public.

The spread of fake news poses a threat to democratic processes, as it can manipulate public opinion, distort electoral outcomes, and erode trust in democratic institutions. Identifying and combating fake news is challenging due to its rapid dissemination, sophisticated manipulation techniques, and the sheer volume of online content.

This fake news classifier works to help mitigate the spread of such information by classifying news as real or fake, based on it's content. This is done by using a Decision Tree Classifier that is trained on a dataset of close to 44,000 data points containing labelled fake news (found in "Fake.csv") and true news(found in "True.csv").

The content of these two files is first cleaned and encoded using a Natural Language Processing pre-processing pipeline where the data is cleaned, removing any values that may cause errors and also normalising it, and is then converted into numeric vectors using TF-IDF vectorisation.

All of this can be used seamlessly using the gradio frontend used within python, which is a popular library for creating interactive user interfaces.

Objectives

1. Develop and deploy a fake news classifier utilizing a Decision Tree Classifier trained on a dataset of approximately 44,000 labeled data points, distinguishing between real and fake news based on content.

2. Implement a Natural Language Processing pre-processing pipeline to clean and encode the content of news articles, removing errors and normalizing the data, followed by TF-IDF vectorization to convert text into numeric vectors.

3. Integrate the fake news classifier seamlessly into a Gradio frontend within Python, providing an interactive user interface for users to input news articles and receive real-time classification results.

Methodology

Steps

1. **Data Collection:**

   Obtain datasets containing labeled true and fake news articles from reliable sources or repositories.

2. **Data Preprocessing:**

- Load the true and fake news datasets.
- Concatenate the datasets and shuffle the rows.
- Clean the text data by removing special characters, URLs, HTML tags, and non-word characters.
- Remove stopwords and lemmatize the words to reduce dimensionality.
- Split the dataset into training and testing sets.

3. **Text Vectorization:**

   Convert the text data into numeric vectors using TF-IDF vectorization.

4. **Model Training:**

   Train a Decision Tree Classifier on the training data.

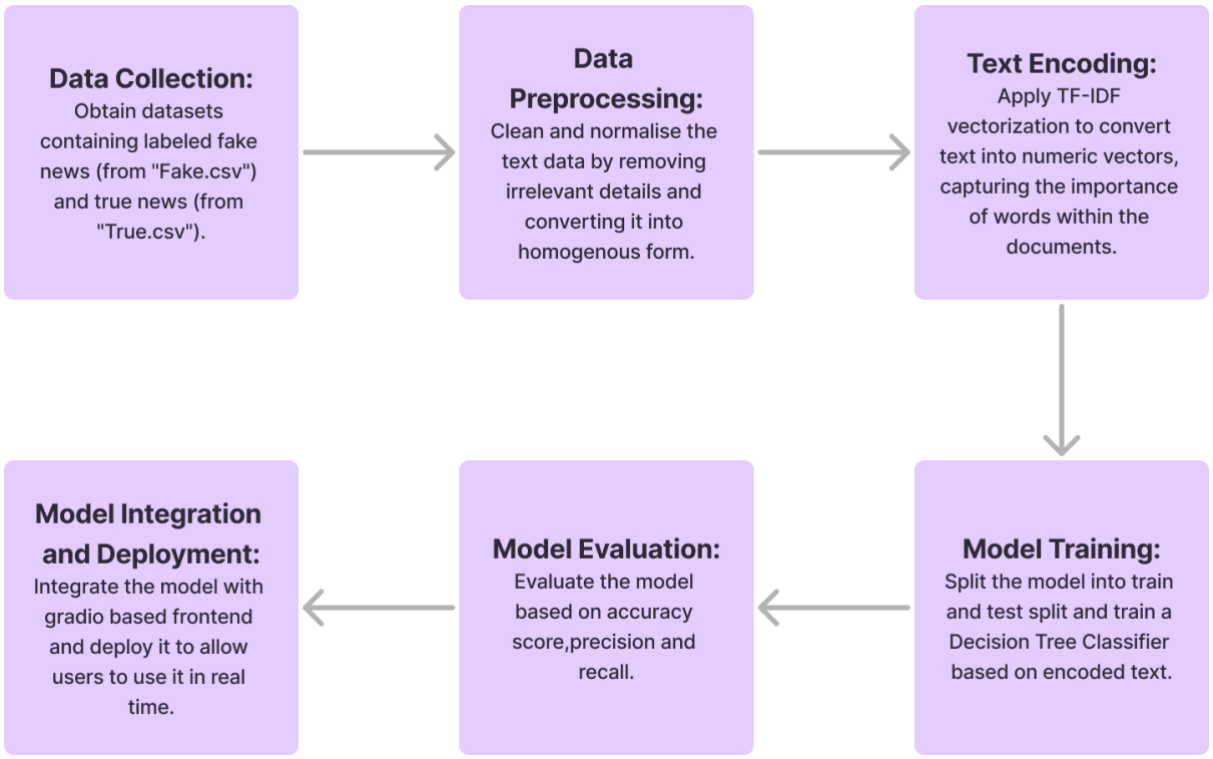
5. **Model Evaluation:**

   Evaluate the model's performance on the testing set using accuracy, confusion matrix, and classification report.

6. **Integration with Gradio Frontend:**

- Create a function to classify news articles using the trained model.
- Implement a Gradio frontend interface to allow users to input news articles.
- Display the classification results to the user in real-time.

Flowchart

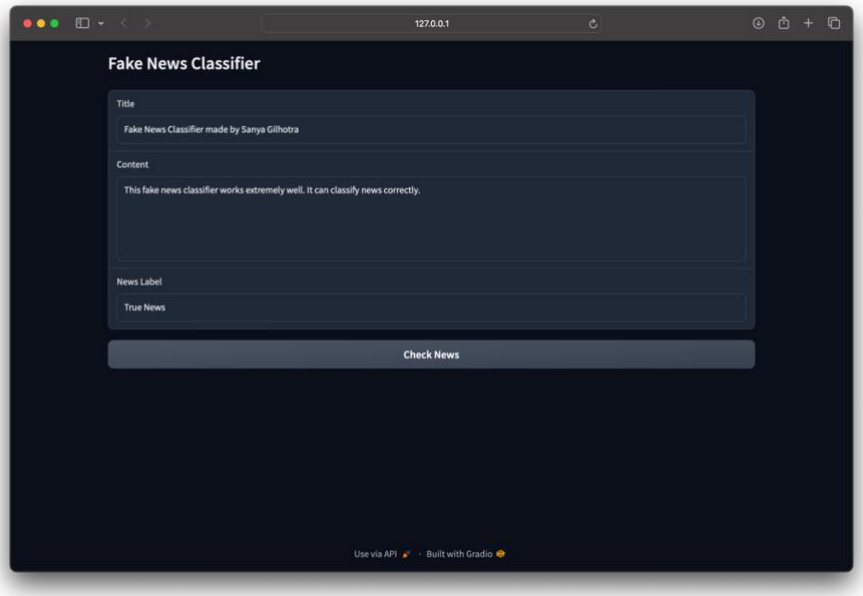


Requirements

- Python libraries: pandas, matplotlib, seaborn, re, string, nltk, sklearn.
- Datasets: "True.csv" and "Fake.csv" containing labeled true and fake news articles.
- Pre-trained NLP models: WordNetLemmatizer from NLTK for lemmatization.
- Gradio library for creating interactive user interfaces.
- Computing resources for data collection, preprocessing, model training, and deploying the Gradio frontend.

Result

Users can add the title and content of any news article or blog to check its credibility and get results instantaneously. This helps people be more aware of the content they consume and also prevent misinformation from causing people to make polarised decisions.



Conclusion

In this project, the aim was to address the challenge of identifying fake news. The process began with the collection and cleaning of datasets containing real and fake news articles. Leveraging machine learning techniques such as TF-IDF vectorization and a Decision Tree Classifier, a system was developed to differentiate between genuine and fabricated news articles with reasonable accuracy.
Following rigorous testing, the effectiveness of the system was confirmed. By integrating it into a user-friendly interface, accessibility was prioritized, allowing individuals to easily verify the authenticity of news articles. This project underscores the role of technology in combating misinformation and emphasizes the importance of critical thinking in navigating today's media landscape.