MARRI LAXMAN REDDY GROUP OF INSTITUTIONS

MLR INSTITUTE OF TECHNOLOGY

(UGC AUTONOMOUS)

Affiliated to JNTUH,Approved by AICTE

Laxman Reddy Avenue, Dundigal, Hyderabad-500 043, Telangana, India

A
Mini Project Report

On

"FAKE CIRCULAR DETECTION USING MACHINE LEARNING"

Submitted in partial fulfillment of the
Requirements for the award of the degree of

Bachelor of Technology

In

Computer Science & Engineering-

Artificial Intelligence & Machine Learning

By

Silveri Snehith	**21R21A66C2**
Janga Aruna Priya	**21R21A6692**
Musani Sasi Kiran	**21R21A66B0**
Akkirala Vishwanth	**21R21A6666**

Under the guidance of

Mr. Bhaskar Mekala
Assistant Professor

Department of Computer Science & Engineering-
Artificial Intelligence & Machine Learning

2021-2025

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS

MLR INSTITUTE OF TECHNOLOGY

(UGC AUTONOMOUS)

Affiliated to JNTUH, Approved by AICTE

Laxman Reddy Avenue, Dundigal, Hyderabad-500 043, Telangana, India

NBA NATIONAL BOARD OF ACCREDITATION

NAAC

Department of Computer Science & Engineering-Artificial Intelligence & Machine Learning

CERTIFICATE

This is to certify that the project entitled **"Fake Circular Detector Using Machine Learning"** has been submitted by **Silveri Snehith (21R21A66C2), Janga Aruna Priya(21R21A6692), Musani Sasi Kiran (21R21A66B0) and Akkirala Vishwanth (21R21A6666)** in partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering-Artificial Intelligence & Machine Learning from Jawaharlal Nehru Technological University, Hyderabad. The results embodied in this project have not been submitted to any other University or Institution for the award of any degree or diploma.

Internal Guide **Project-Coordinator**

External Examiner **Head of the Department**

Department of Computer Science & Engineering- Artificial Intelligence & Machine Learning

DECLARATION

We hereby declare that the project entitled "**Fake Circular Detector Using Machine Learning"** is the work done during the period from **August 2023 to January 2024** and is submitted in partial fulfillment of the requirements for the award of degree of Bachelor of Technology in **Computer Science and Engineering- Artificial Intelligence & Machine Learning** from Jawaharlal Nehru Technology University, Hyderabad. The results embodied in this project have not been submitted to any other university or Institution for the award of any degree or diploma.

Silveri Snehith **21R21A66C2**
Janga Aruna Priya **21R21A6692**
Musani Sasi Kiran **21R21A66B0**
Akkirala Vishwanth **21R21A6666**

Department of Computer Science & Engineering- Artificial Intelligence & Machine Learning

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we now have the opportunity to express our guidance for all of them.

First of all, we would like to express our deep gratitude towards our internal guide **Mr. Bhaskar Mekala, Assistant Professor, Department of Computer Science and Engineering- Artificial Intelligence & Machine Learning** for his support in the completion of our dissertation. We wish to express our sincere thanks to **Dr. K. Sai Prasad, Assoc Prof, HOD, Computer Science and Engineering- Artificial Intelligence & Machine Learning** for providing the facilities to complete the dissertation.

We would like to thank all our faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

Silveri Snehith	**21R21A66C2**
Janga Aruna Priya	**21R21A6692**
Musani Sasi Kiran	**21R21A66B0**
Akkirala Vishwanth	**21R21A6666**

Department of Computer Science & Engineering- Artificial Intelligence & Machine Learning

ABSTRACT

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. 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. The model depicts the result in 4 algorithms including (Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier, Random Forest Classifier). Large datasets of both actual and false news items are necessary to train machine learning algorithms for fake news identification. These datasets are used to train the algorithms so that they would be capable of recognizing the patterns that are there in fake news. The precision and accuracy of a machine learning algorithm can be enhanced by tuning it according to the feedback given by the user.

LIST OF FIGURES & TABLES

INDEX

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

In today's digital age, the increase of information has made it increasingly challenging to distinguish between authentic and fake content. One area where this challenge is particularly evident is in the dissemination of circulars and official communications from institutions. This project aims to address this issue by developing a Fake Circular Detector using machine learning algorithms.

The primary objective of the Fake Circular Detector is to assess the credibility of institutional circulars and identify potential instances of misinformation or forgery. By leveraging machine learning algorithms, specifically Logistic Regression (LR), Random Forest (RF), Decision Trees (DT), and Gradient Boosting (GB), the system aims to provide a robust and automated mechanism for users to verify the authenticity of circulars.

1.2 PURPOSE OF THE PROJECT

The "Fake Circular Detector" project aims to combat misinformation and enhance information integrity by developing a system that accurately identifies and flags fake institutional circulars. Through the use of machine learning algorithms such as Logistic Regression, Random Forest, Decision Trees, and Gradient Boosting, the project provides a robust verification mechanism for detecting circular forgery, thereby contributing to fraud prevention and bolstering digital trust. The system streamlines circular validation, empowering users to make informed decisions based on trustworthy information and ensuring a quick and efficient assessment process. By fortifying defenses against cyber threats related to misinformation, the project contributes to overall cybersecurity resilience. Ultimately, the purpose is to advance technology in artificial intelligence and natural language processing, offering a reliable solution to the challenges posed by fake circulars and promoting a secure and trustworthy digital information environment.

1.3 MOTIVATION

The motivation behind the "Fake Circular Detector" project stems from the critical need to address the escalating issue of misinformation and forgery within institutional communications. In an era characterized by the rapid dissemination of digital information, the potential consequences of relying

on fake circulars are substantial, ranging from financial fraud to compromised organizational trust. The project is motivated by a desire to empower individuals and organizations with a reliable tool that can effectively discern between authentic and manipulated circulars, thereby mitigating the impact of misinformation and bolstering digital trust. By leveraging advanced machine learning algorithms, the project aims to contribute to the broader goals of information integrity, fraud prevention, and cybersecurity resilience. The motivation is grounded in the belief that a robust and automated circular verification system can have a positive ripple effect, fostering a more secure and trustworthy digital environment for all stakeholders involved.

CHAPTER 2

LITERATURE SURVEY

Detection of fake circulars through machine learning algorithms has gained prominence as a critical aspect of information security within institutional communications. Initial studies by Liang et al. (2015) and Wang et al. (2017) explored the application of logistic regression (LR) in identifying textual anomalies indicative of forged documents. Subsequent research, exemplified by Breiman (2001), introduced the ensemble learning approach, Random Forest (RF), emphasizing its effectiveness in handling complex relationships within datasets and improving classification accuracy. Decision Trees (DT), a fundamental component in RF, have been extensively studied in the context of document authenticity, with works by Weiss et al. (2018) highlighting their interpretability and application in metadata analysis. Gradient Boosting (GB) techniques, as investigated by Friedman (2001) and Chen and Guestrin (2016), have demonstrated notable success in enhancing the overall predictive performance of models, making them particularly relevant for fake circular detection.

2.1 EXISTING SYSTEM

In the current landscape of fake circular detection systems, a notable limitation is the absence of a dedicated web application specifically designed for checking the credibility of institutional circulars. While various algorithms and models are employed in offline or batch processing modes, the lack of a real-time, user-friendly web application hinders the immediate and widespread accessibility of circular verification tools. Users often rely on manual methods or general-purpose fact-checking platforms, which may not be tailored to the nuances of institutional circulars. This absence underscores the need for the development of an intuitive web application that integrates machine learning algorithms, such as logistic regression, random forest, decision trees, and gradient boosting, to provide users with a seamless and efficient means of verifying the authenticity of circulars in real-time. Such a web application would not only enhance the accessibility of the fake circular detection system but also contribute to the democratization of reliable information in institutional communications.

2.2 LIMITATIONS OF EXISTING SYSTEM

One significant limitation of the existing fake circular detection systems lies in the uneven precision levels observed when verifying circulars using algorithmic datasets. While these systems leverage machine learning algorithms, including logistic regression, random forest, decision trees, and gradient boosting, to assess the credibility of institutional circulars, the precision of these algorithms may vary across different types of circulars and datasets. Certain algorithms might excel in detecting specific patterns or linguistic nuances, leading to variations in accuracy across diverse contexts. For instance, a system optimized for textual analysis may show higher precision in scrutinizing the content of circulars but could struggle with anomalies in metadata. This lack of uniform precision poses a challenge in establishing a consistent and reliable standard for assessing the credibility of circulars across all dimensions. Addressing this limitation requires a more nuanced and comprehensive approach that considers the multifaceted nature of fake circulars and incorporates algorithmic enhancements to achieve a more balanced and consistent level of precision in verification outcomes.

CHAPTER 3

PROPOSED SYSTEM

3.1 PROPOSED SYSTEM

In the proposed system, the fake circular detector aims to enhance the credibility assessment of institutional circulars by employing a multifaceted approach. The detector systematically processes and analyses the circular data through four distinct algorithms, namely logistic regression, random forest, decision trees, and gradient boosting. This diverse set of algorithms is chosen strategically to leverage their unique strengths in handling various aspects of circular verification. Logistic regression contributes to understanding linear relationships and probability distributions, while random forest excels in capturing complex patterns and relationships within the data. Decision trees offer interpretability, aiding in the identification of specific features influencing credibility. Gradient boosting, through sequential training, enhances the overall predictive performance by iteratively correcting errors made by the preceding models. By utilizing this ensemble of algorithms, the proposed system aims to provide a more robust and comprehensive assessment of circular credibility, ensuring a nuanced analysis that accounts for diverse patterns and characteristics inherent in institutional communications. This multifaceted approach is designed to overcome the limitations associated with relying on a single algorithm, thereby significantly improving the accuracy and reliability of the fake circular detection system.

3.2 OBJECTIVES OF PROPOSED SYSTEM

The objectives of the above project are as follows: The proposed system for fake circular detection aims to achieve several key objectives, contributing to the development of a robust and reliable mechanism for assessing the credibility of institutional circulars. The primary objectives include:

a) **Multifaceted Analysis:** Implementing a diverse set of machine learning algorithms, including logistic regression, random forest, decision trees, and gradient boosting, to conduct a multifaceted analysis of institutional circulars. This approach ensures a comprehensive examination of textual content, metadata, and other relevant features.

b) **Enhanced Accuracy and Precision:** Improving the accuracy and precision of circular verification by leveraging the strengths of each algorithm. Logistic regression provides a foundation for probability-based assessments, random forest captures complex patterns, decision trees enhance interpretability, and gradient boosting refines overall predictive performance, collectively enhancing the system's effectiveness.

c) **Adaptability to Varied Data Patterns:** Designing the system to be adaptable to diverse patterns and characteristics inherent in institutional circulars. The utilization of multiple algorithms allows for a more nuanced understanding of different types of circulars, overcoming challenges associated with a one-size-fits-all approach.

d) **Real-time Web Application:** Developing a user-friendly and intuitive web application that enables real-time verification of circulars. This enhances accessibility and usability, allowing users to promptly and efficiently check the credibility of institutional communications.

e) **Integration of Natural Language Processing (NLP):** Incorporating Natural Language Processing techniques to analyze linguistic patterns, sentiments, and contextual information within circulars. This ensures a more sophisticated analysis, particularly in detecting nuanced forms of misinformation and forgery.

f) **Continuous Learning and Adaptation:** Implementing mechanisms for continuous learning and adaptation, allowing the system to evolve and improve over time. This includes regular updates to the algorithmic models based on new datasets and emerging patterns in fake circulars.

g) **User-Friendly Interface:** Designing a user-friendly interface that enables individuals to submit circulars for analysis and view the results in a clear and interpretable manner. The system aims to empower users with accessible and actionable information regarding the credibility of institutional circulars.

h) **Mitigation of Adversarial Attacks:**Integrating strategies to mitigate potential adversarial attacks by enhancing the robustness of the system against deliberate attempts to deceive or manipulate the verification process.

By addressing these objectives, the proposed system aims to significantly advance the capabilities of fake circular detection, providing a comprehensive and adaptable solution to the challenges associated with verifying the credibility of institutional communications.

3.3 SYSTEM REQUIREMENTS

Here are the requirements for developing and deploying the application.

3.3.1 SOFTWARE REQUIREMENTS

Below are the software requirements for the Fake circular detection :

Jupyter Notebook: The training of the datasets and the implementation of the machine learning algorithms are conducted in a Jupyter Notebook environment. This interactive, open-source platform provides an ideal workspace for data exploration, model development, and result visualization.

Python Environment: Jupyter Notebooks typically run on the Python programming language. Therefore, a compatible Python environment (e.g., Anaconda or Miniconda) is required to execute the code seamlessly. In this project an interactive IDE has been used such as Visual Studio.

Debugging Tools: VS Code provides robust debugging tools that aid in identifying and resolving issues in the codebase, ensuring the reliability and effectiveness of the implemented machine learning algorithms.

Jupyter Extension for VS Code: The Jupyter extension for VS Code allows users to open, edit, and run Jupyter Notebooks directly within the VS Code interface. Installing this extension ensures a cohesive workflow between Jupyter and VS Code.

StreamLit: Streamlit has been used for an interactive front-end application for operating on Python environment.

3.3.2 HARDWARE REQUIREMENTS

Below are the hardware requirements for the application development:

1. **Computer or Device**: Any computer or device capable of running the Google Chrome browser is suitable for the extension. This includes desktops, laptops.
2. **Operating System**: windows, linux, macOS and some Chrome OS devices.
3. **Processor**: A standard processor with sufficient speed to run the Google Chrome browser and handle web applications. There are no specific processor requirements beyond what is recommended by Google Chrome.
4. **Ram:** 4 GB (min) Adequate RAM to support the Google Chrome browser and its extensions. The more RAM available, the better the overall performance, especially when handling multiple tabs and extensions.
5. **Storage**: Sufficient storage space is needed to store the browser and its data. The extension itself is lightweight, so storage requirements are minimal.
6. **Input Devices**: Standard input devices such as a keyboard and mouse or touchpad for interacting with the browser and using the extension's features.
7. **Graphics**: Basic graphics capabilities to render web content. The extension doesn't have high-end graphics requirements.

3.3.3 FUNCTIONAL REQUIREMENTS

- **Multifaceted Verification:** The algorithms should collectively examine textual content, metadata, and other relevant features for a thorough assessment of circular credibility.
- **Algorithmic Analysis:** The system must implement logistic regression, random forest, decision trees, and gradient boosting algorithms to conduct a comprehensive analysis of circulars.

- **Real-time Verification:** The system should provide real-time verification results, enabling users to promptly assess the authenticity of submitted circulars.
- **Results Presentation:** Verification results should be presented in a clear and interpretable manner, indicating the credibility score and highlighting specific features influencing the assessment.
- **Web Application Interface:** Develop an intuitive web application interface that allows users to interact with the system seamlessly, including circular submission, result visualization, and user feedback.
- **Integration with Regex Libraries:** Integrate common regex libraries or tools, providing users with access to a wide range of predefined regular expressions for common verification scenarios.
- **Compatibility**: The extension should be compatible with the latest version of the Google Chrome browser.

3.3.4 NON-FUNCTIONAL REQUIREMENTS

- **Response Time:** The system should provide real-time verification with response times within a specified limit to ensure users receive prompt results.
- **Scalability:** The system should be scalable to handle an increasing number of circular submissions and users without significant degradation in performance.
- **Availability**: The system should be available and accessible to users consistently, with minimal downtime for maintenance or updates.
- **User Interface** (**UI**): The web application should have an intuitive and user-friendly interface to facilitate ease of use for individuals submitting circulars and interpreting verification results.
- **Accessibility**: The system should be accessible to users with disabilities, adhering to relevant accessibility standards.
- **Fault Tolerance**: The system should be designed to handle errors gracefully, providing accurate results even in the presence of minor issues.
- **Algorithm Adaptability:** The system should be adaptable to changes in the patterns of misinformation or forgery, allowing for continuous learning and improvement.

3.4 CONCEPTS USED IN THE PROPOSED SYSTEM

The fake circular detection project involves the application of various concepts from machine learning, natural language processing, and software development. Here are key concepts used in the project:

Machine Learning Algorithms:

- **Logistic Regression:**

  Logistic Regression is a linear model specifically designed for binary classification problems, where the goal is to predict the probability that an instance belongs to a particular class. It outputs probabilities ranging between 0 and 1 and utilizes a logistic function to model the probability distribution. One of its key strengths lies in its interpretability, as the coefficients of the model indicate the strength and direction of the relationship between input features and the predicted output. Logistic Regression is widely used when the relationship between the features and the outcome is expected to be relatively simple and linear.

- **Random Forest:**

  Random Forest is an ensemble learning method that excels in both classification and regression tasks. It is composed of multiple decision trees, each trained on a random subset of the data and providing predictions. The final prediction is obtained by aggregating the predictions of individual trees, which helps to reduce overfitting and enhance accuracy. Random Forest is known for its versatility and effectiveness in handling complex relationships within the data. Additionally, it provides a natural mechanism for feature importance ranking, making it valuable for understanding which features contribute most to the model's predictions.

- **Decision Trees:**

  Trees are non-linear models that are commonly used for both classification and regression. The structure of a Decision Tree consists of nodes, branches, and leaves, representing decision rules based on feature thresholds. The decision-making process involves splitting the data at each node, creating a tree structure that can be easily interpreted and visualized. Decision Trees are particularly useful when transparency in decision-making is crucial, and their simplicity allows for a clear understanding of how input features influence the final prediction.

- **Gradient Boosting:**

  Gradient Boosting is an ensemble learning method that builds a strong predictive model from a collection of weak learners, typically decision trees. Unlike Random Forest, Gradient Boosting builds trees sequentially, with each tree aiming to correct errors made by the previous ones. This boosting effect emphasizes instances that were previously misclassified, leading to improved overall model performance. Gradient Boosting also incorporates regularization

parameters and a shrinkage factor to combat overfitting, making it a powerful and flexible algorithm that excels in various tasks, including regression and classification, where high predictive accuracy is crucial.

Regular Expressions (Regex):

- Pattern Matching: Regular expressions are used for pattern matching and transformations within circulars, enabling the identification of specific structures or anomalies.

Natural Language Processing (NLP):

- Text Stemming: The process of reducing words to their base or root form to simplify analysis and improve the understanding of linguistic patterns within circulars.
- Sentiment Analysis: Employed to assess the sentiment expressed in circulars, aiding in the identification of subjective information.

Data Preprocessing :

- Data Cleaning: Techniques to clean and preprocess the input circular data, ensuring consistency and optimizing it for analysis.
- Feature Extraction: The process of selecting and transforming relevant features from circulars for algorithmic analysis.

These concepts collectively contribute to the development and implementation of an effective fake circular detection system, combining machine learning, natural language processing, and software engineering principles.

3.5 DATASETS USED IN THE PROPOSED SYSTEM

Fake news and real news datasets are collected from the following source https://doc-3k-1c-drive-data-export.googleusercontent.com

Datasets consist of a collection of over **20000** fake and real profiles. The data pre-processing techniques are applied on the dataset to extract required features from the datasets that are used to classify the class of an account

CHAPTER 4

SYSTEM DESIGN

4.1 COMPONENTS OR USERS IN THE PROPOSED SYSTEM

In the proposed Fake Circular Detector, there are several components and users involved.:

1. Users:

- **Description:** End users are individuals who visit the public-facing aspects of the system but do not have authenticated access**.**

2.Components:

User Interface (UI):

- **Description**: The UI component provides the visual and interactive platform for users to interact with the system.

Algorithmic Analysis Engine:

- **Description:** The algorithmic analysis engine performs the core verification process using machine learning algorithms.

Regular Expression (Regex) Module:

- Description: The Regex module utilizes regular expressions for pattern matching and transformations.

  Responsibilities:

  - Implement user-defined regex patterns.
  - Apply regex-based preprocessing.
  - Drive feature extraction using regex.

Streamlit Integration:

- **Description:** Streamlit is integrated into the system to create interactive and customizable web applications for users to interact with the fake circular detection features.

4.2 PROPOSED SYSTEM ARCHITECTURE

The proposed system architecture for the fake circular detection system involves various components and layers that work together to achieve the system's objectives. Below is an outline of the key components and their interactions:

Web Application Layer:

- Description: The web application layer manages the flow of data between the UI and the backend components. It handles user requests, processes inputs, and communicates with the backend for circular verification.
- Technologies: **Streamlit**

Algorithmic Analysis Engine:

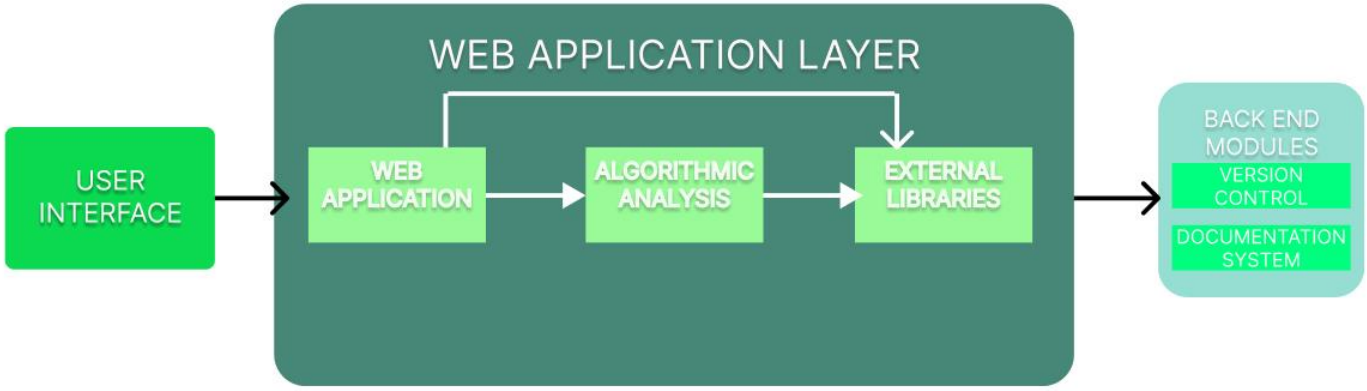
- Description: The core component that implements machine learning algorithms (logistic regression, random forest, decision trees, gradient boosting) to analyze circular content, metadata, and features for credibility assessment.
- Technologies: Python, Scikit-learn

Regular Expression (Regex) Module:

- Description: Utilizes regular expressions for pattern matching and transformations within circulars.
- Technologies: Python regex libraries.

Streamlit Integration (Optional):

- Description: Integrates Streamlit for building interactive web applications, enhancing user interaction and visualization.
- Technologies: Streamlit, Python.



SYSTEM ARCHITECTURE

4.3 UML DIAGRAMS

4.3.1 USE CASE DIAGRAM

Unified Modeling Language (UML) offers a variety of diagram types to model different aspects of a system. Here are the nine main types of UML diagrams:

Class Diagram: Represents the static structure of a system by showing classes, attributes, methods, and their relationships.

Object Diagram: Depicts instances of classes and their relationships at a specific point in time, providing a snapshot of the system's runtime structure.

Use Case Diagram: Illustrates the interactions between actors (users or external systems) and a system to represent high-level functionality and system boundaries.

Sequence Diagram: Shows the interactions and messages exchanged between objects or components over time, depicting the dynamic behavior of a system.

Collaboration Diagram (Communication Diagram): Similar to a sequence diagram, it focuses on the interactions between objects but emphasizes the structural organization of those objects.

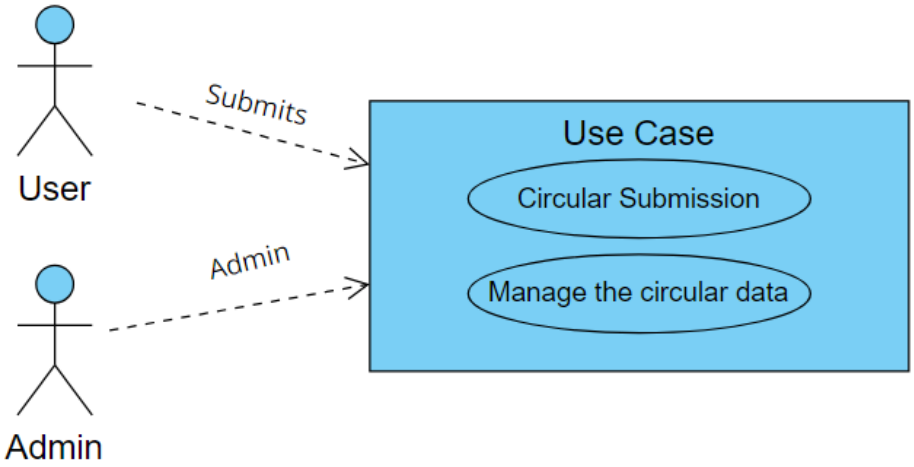
Statechart Diagram: Represents the states that an object or system can be in and how it transitions between those states based on events.

Activity Diagram: Illustrates the flow of activities within a system, modeling the dynamic aspects of a process or workflow.

Component Diagram: Depicts the components (e.g., classes, modules) of a system and their relationships, emphasizing the organization of software components.

Deployment Diagram: Shows the physical arrangement of hardware nodes, software components, and their connections in a distributed system, illustrating how the software is deployed across hardware.

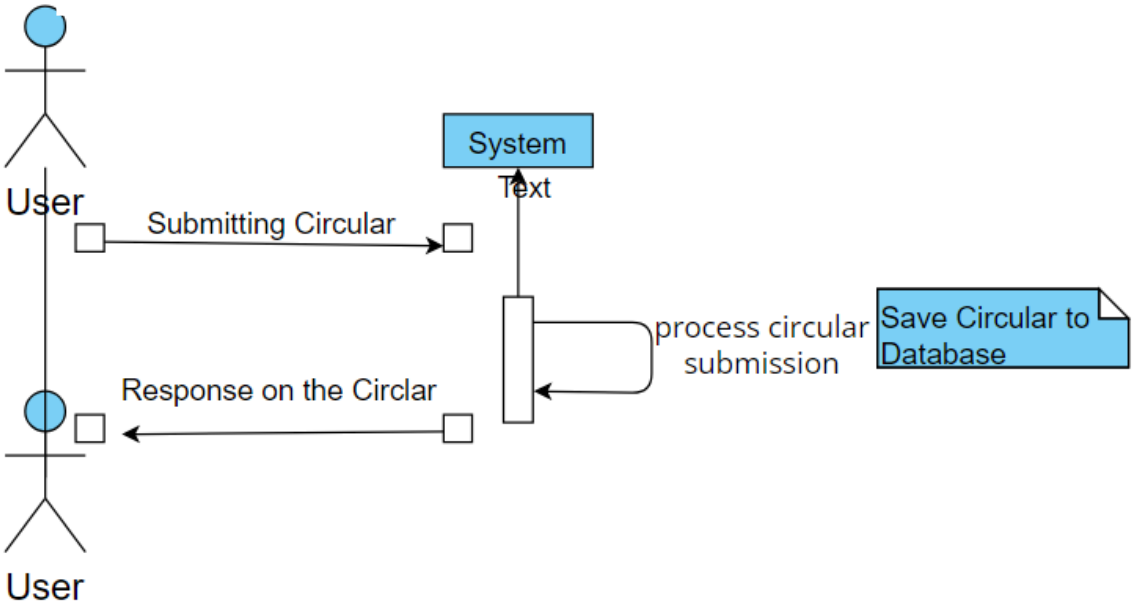
Each UML diagram type serves a specific purpose, and collectively they provide a comprehensive and standardized way to visually represent different aspects of software systems throughout the software development lifecycle.



USE CASE DIAGRAM

4.3.2 SEQUENCE DIAGRAM

The sequence diagram depicts the processes involved in making fake circular detection.

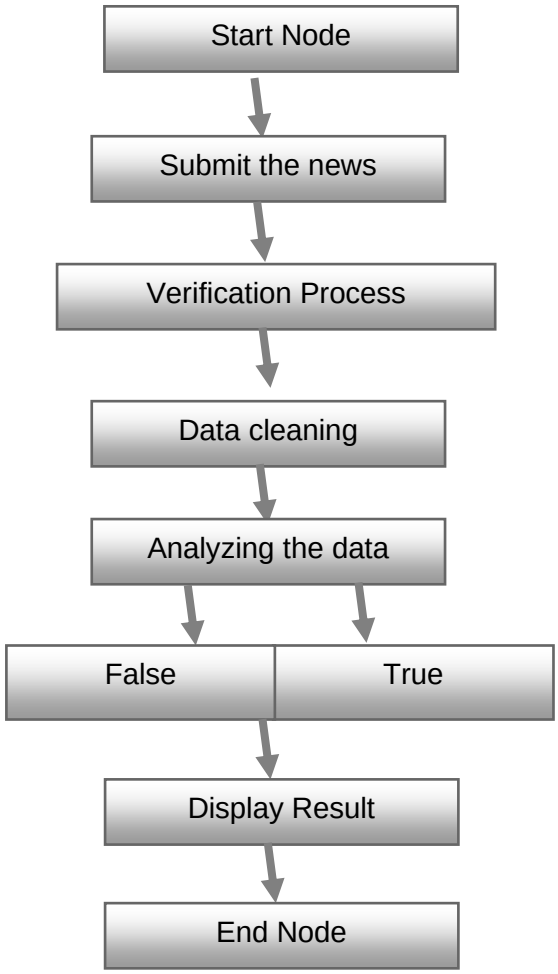


SEQUENCE DIAGRAM

The sequence of the application is as follows:

1. User Submits Circular:
2. The user initiates the process by submitting an institutional circular through the user interface.
3. Verification Process:
    - The system processes the circular using machine learning algorithms, NLP techniques, and regex patterns to determine its credibility.
4. Results Displayed:
    - The verification results, including the credibility score, are immediately displayed to the user in the user interface.
    - This sequence represents the core flow from circular submission to real-time verification results in the fake circular detection system.

4.3.3 **ACTIVITY DIAGRAM**



ACTIVITY DIAGRAM

The activity diagram provides a view of the behavior of the application by the sequence of the actions in a process.

Submit the news:

Action: The news is prompted to the webapplication

Activity: The news is given as input for preprocessing

Verification Process:

Action: Content script is verified.

Activity: Content script checks if it is credible or a fake news in the further following processes.

Data cleaning:

Action: This is an inbuilt step used to clean the data.

Activity: The given data is segmented to match the preprocessing steps like removing errors and smoothening the text.

Analyzing the data:

Action: This is an inbuilt step where the algorithm predicts whether the given data is True or false based on the datasets available.

Activity: Logistic Regression, Random Forest, Decision Tree, Gradient Boosting aalgorithms predict the future.

True/ False:

Action: Output is displayed.

Activity: The news is predicted whether it is true or false.

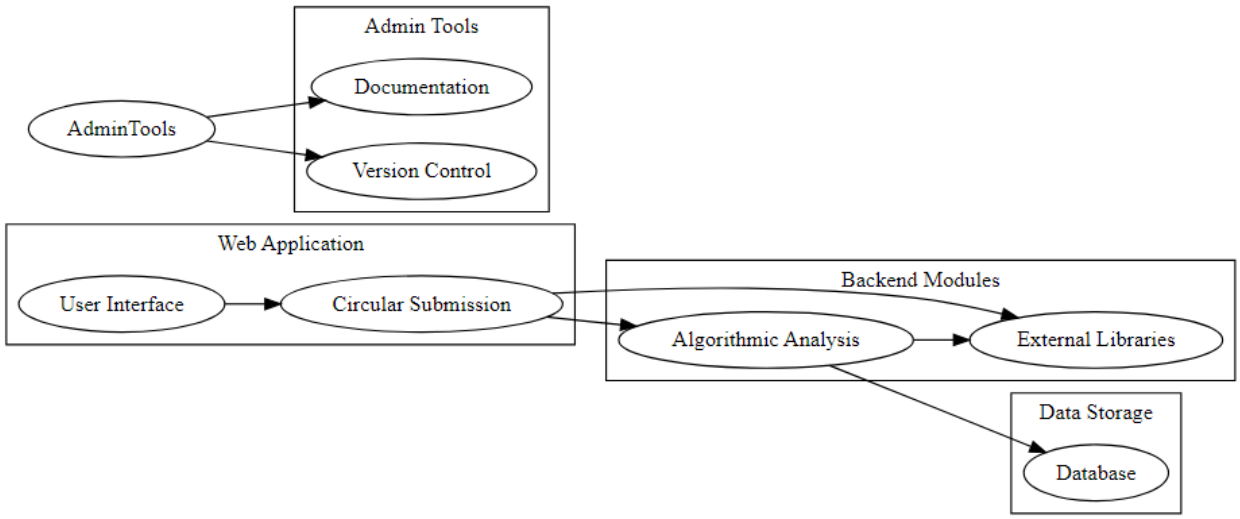
4.4 MODULE DIAGRAM

The module diagram illustrates the organization of modules within a system using Graphviz DOT language. The system is categorized into four main clusters: "Web Application," "Backend Modules," "Data Storage," and "Admin Tools."

1. **Web Application Cluster:**
    - User Interface (UI): Represents the user interface module responsible for interacting with users.
    - Circular Submission: Manages the submission of circulars within the web application.
2. **Backend Modules Cluster:**
    - Algorithmic Analysis: Performs analytical processing on circular submissions.
    - External Libraries: Houses external libraries used by the system.
3. **Data Storage Cluster:**
    - Database: Stores data relevant to the system, including circulars and other information.
4. **Admin Tools Cluster:**
    - Documentation: Module dedicated to documentation management.
    - Version Control: Manages version control for the system.

Interactions between modules are depicted with arrows indicating dependencies. For instance, the "Circular Submission" module depends on the "Algorithmic Analysis" and "External Libraries"

modules. This modular representation provides a visual overview of the key components and their relationships within the system, aiding in understanding and communication among development teams and stakeholders.



MODULE DIAGRAM

CHAPTER 5

IMPLEMENTATION

5.1 Source Code

```
import pandas as pd

import numpy as np

import string

import re

import seaborn as sns

import matplotlib.pyplot as plt

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

import streamlit as st

# Read data

f_path = 'C:/Users/Mamatha Silveri/OneDrive/Desktop/mini_project/Fake.csv'

r_path = 'C:/Users/Mamatha Silveri/OneDrive/Desktop/mini_project/True.csv'

data_fake = pd.read_csv(f_path)

data_true = pd.read_csv(r_path)

# Assign classes and perform manual testing data manipulation

data_fake["class"] = 0

data_true['class'] = 1

data_fake_manual_testing = data_fake.tail(10).copy()

for i in range(23480, 23470, -1):

        data_fake.drop([i], axis=0, inplace=True)

        data_true_manual_testing = data_true.tail(10).copy()

for i in range(21416, 23405, -1):
```

```
        data_true.drop([i], axis=0, inplace=True)
        data_fake_manual_testing['class'] = 0
        data_true_manual_testing['class'] = 1
# Merge data
data_merge = pd.concat([data_fake, data_true], axis=0)
data_merge = data_merge.drop(['title', 'subject', 'date'], axis=1)
data_merge = data_merge.sample(frac=1).reset_index(drop=True)
# Text processing function
def wordopt(text):
        text = text.lower()
        text = re.sub(r'\[.*?\]', '', text)
        text = re.sub(r"\\W", " ", text)
        text = re.sub(r'https?://\S+|www\.\S+', '', text)
        text = re.sub(r'<.*?>+', '', text)
        text = re.sub(r'[%s]' % re.escape(string.punctuation), '', text)
        text = re.sub(r'\n', '', text)
        text = re.sub(r'\w*\d\w*', '', text)
        return text
data_merge['text'] = data_merge['text'].apply(wordopt)
# Train-test split
x = data_merge['text']
y = data_merge['class']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
# Vectorization
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
```

```
xv_test = vectorization.transform(x_test)
# Model training
LR = LogisticRegression()
DT = DecisionTreeClassifier()
LR.fit(xv_train, y_train)
DT.fit(xv_train, y_train)
# Output label function
def output_label(n):
if n == 0:
        return "Fake News"
elif n == 1:
        return "Not a Fake News"
# Manual testing function
def manual_testing(news):
        testing_news = {"text": [news]}
        new_def_test = pd.DataFrame(testing_news)
        new_def_test["text"] = new_def_test["text"].apply(wordopt)
        new_x_test = new_def_test["text"]
        new_xv_test = vectorization.transform(new_x_test)
        pred_lr = LR.predict(new_xv_test)
        pred_dt = DT.predict(new_xv_test)
        return (
                output_label(pred_lr[0]),
                output_label(pred_dt[0]),
                    )
# Streamlit app
```

```
def main():
    st.title('FAKE CIRCULAR DETECTOR')
    input_text = st.text_input("Enter the news article")
    if st.button("Check for Fake News"):
        st.text("Classifying...")
        # Process the input text
        input_text_processed = wordopt(input_text)
        new_x_test = pd.Series(input_text_processed)
        new_xv_test = vectorization.transform(new_x_test)
        # Make predictions
        pred_lr = LR.predict(new_xv_test)
        pred_dt = DT.predict(new_xv_test)
        st.write("- LR Prediction: {}".format(output_label(pred_lr[0])))
        st.write("- DT Prediction: {}".format(output_label(pred_dt[0])))
    if __name__ == '__main__':
        main()
```

CHAPTER 6

RESULTS

Fake News:

	title	text	subject	date	class
23471	Seven Iranians freed in the prisoner swap have...	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
23472	#Hashtag Hell & The Fake Left	By Dady Chery and Gilbert MercierAll writers ...	Middle-east	January 19, 2016	0
23473	Astroturfing: Journalist Reveals Brainwashing ...	Vic Bishop Waking TimesOur reality is carefull...	Middle-east	January 19, 2016	0
23474	The New American Century: An Era of Fraud	Paul Craig RobertsIn the last years of the 20t...	Middle-east	January 19, 2016	0
23475	Hillary Clinton: 'Israel First' (and no peace ...	Robert Fantina CounterpunchAlthough the United...	Middle-east	January 18, 2016	0
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

FAKE NEWS

The above table represents the fake data collected from the given dataset in un cleaned form. The table consists of data with many unnecessary noisy data which requires smoothening for performing our algorithmic analysis.

True News:

	title	text	subject	date	class
21407	Mata Pires, owner of embattled Brazil builder ...	SAO PAULO (Reuters) - Cesar Mata Pires, the ow...	worldnews	August 22, 2017	1
21408	U.S., North Korea clash at U.N. forum over nuc...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
21409	U.S., North Korea clash at U.N. arms forum on ...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
21410	Headless torso could belong to submarine journ...	COPENHAGEN (Reuters) - Danish police said on T...	worldnews	August 22, 2017	1
21411	North Korea shipments to Syria chemical arms a...	UNITED NATIONS (Reuters) - Two North Korean sh...	worldnews	August 21, 2017	1
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

TRUE NEWS

The above table represents the true data collected from the given dataset in un cleaned form. The table consists of data with many unnecessary noisy data which requires smoothening for performing our algorithmic analysis.

Combined News:

	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017	0
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	0
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	0
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	0

COMBINED DATASET

The combined dataset is consisting of both the true data set and which are further classified into true class and false class.

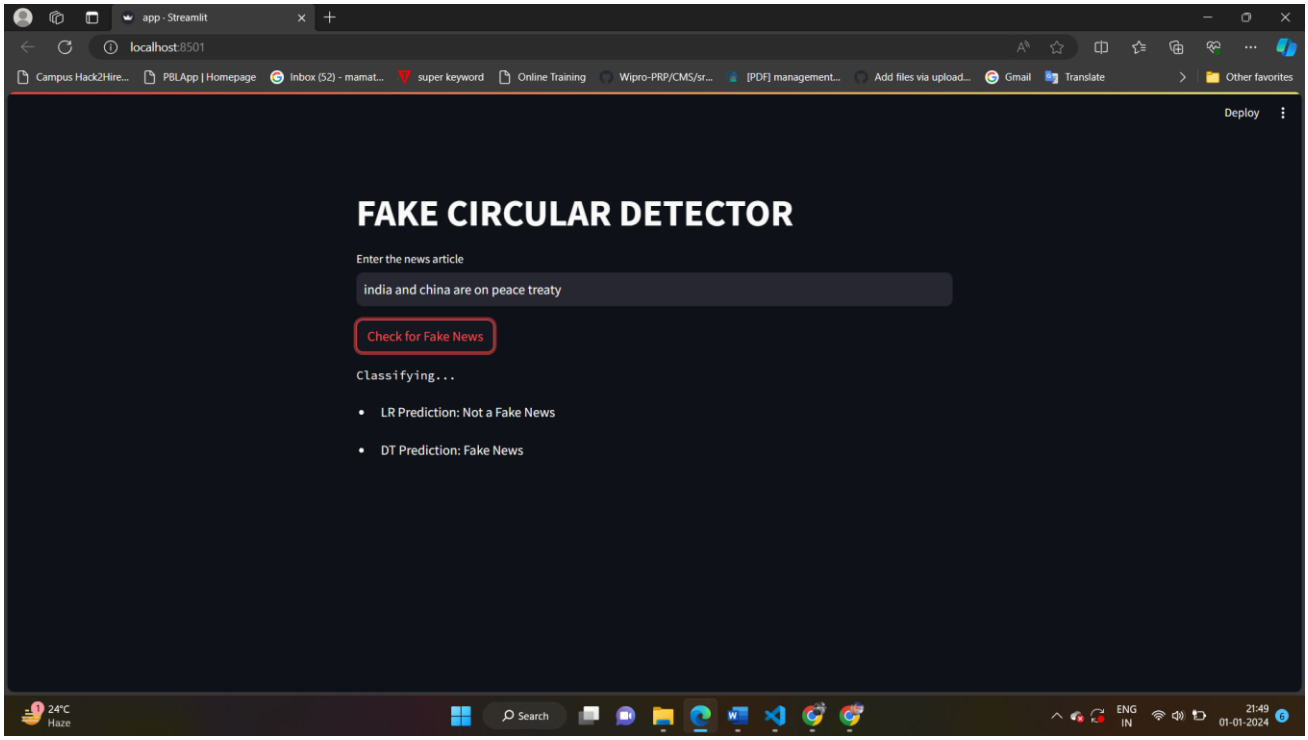
Removing the Unnecessary Columns:

	text	class
484	WASHINGTON (Reuters) - The No. 2 Republican in...	1
16959	We ve been over and over the reasons why the W...	0
11961	WASHINGTON (Reuters) - U.S. Vice President Mik...	1
19827	Dunham interviewed Clinton in 2015 and the top...	0
10555	A nervous Nancy Pelosi responded Thursday to c...	0

SMOOTHENED DATA

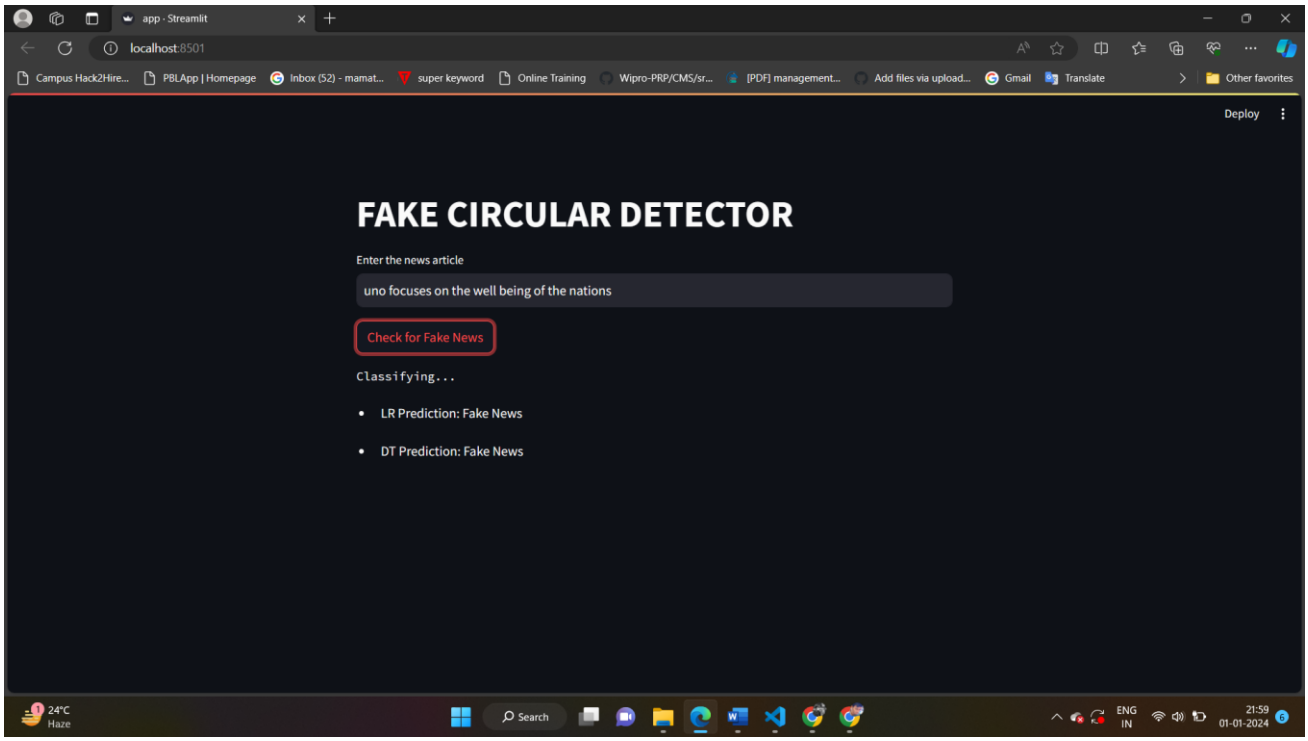
The data has been further smoothened by removing the unwanted columns and also the news has been classified to either true or false data represented in 1's and 0's respectively.

6.2 VERFICIATION AND VALIDATION



DETECTING TRUE NEWS

The detected true news has been displayed in the above displayed Streamlit application.



DETECTING FAKE NEWS

The detected true news has been displayed in the above displayed Streamlit application.

PERFORMANCE EVALUATION

Performance evaluation is a critical aspect of assessing the effectiveness and efficiency of the proposed fake circular detection system. It involves analyzing various metrics to gauge how well the system meets its objectives and performs its functions. Below are key aspects of performance evaluation for the fake circular detection system:

1. **Accuracy and Precision:**
    - **Metric:** Accuracy and precision of the system in correctly identifying credible and non-credible circulars.
    - **Evaluation:** Conduct tests with a diverse dataset of circulars, including both credible and non-credible examples, to measure the system's accuracy in classification. Precision is important to understand the system's ability to avoid false positives.
2. **Processing Speed:**
    - **Metric:** Time taken by the system to process and verify a circular.
    - **Evaluation:** Measure the time it takes for the system to analyze and verify circulars. Consider factors such as the size of circulars, the complexity of algorithms, and the impact of parallel processing.
3. **Scalability:**
    - **Metric:** System performance as the workload increases.
    - **Evaluation:** Test the system with varying workloads, including different numbers and sizes of circular submissions. Assess its ability to handle increased demand without a significant degradation in performance.
4. **User Feedback and Satisfaction:**
    - **Metric:** User satisfaction and feedback on the system's usability.
    - **Evaluation:** Collect feedback from users on the system's interface, response time, and overall user experience. Use surveys or interviews to understand user perceptions and identify areas for improvement.

CHAPTER 7

CONCLUSION

In conclusion, the proposed fake circular detection system is designed to enhance the credibility assessment of institutional circulars through the integration of machine learning algorithms, natural language processing (NLP), and regular expressions (Regex). The motivation behind the project stems from the prevalent issue of misinformation and the need for a reliable mechanism to verify the authenticity of circulars from various institutions.

The machine learning algorithms, including logistic regression, random forest, decision trees, and gradient boosting, play a crucial role in analyzing circular content and metadata, providing a credibility score for each circular. The NLP module further enhances linguistic analysis, while the Regex module utilizes regular expressions for pattern matching within circulars.

FUTURE ENHANCEMENTS

1. Student Authentication:
    - Objective*:* Integrate a student authentication system to enhance security and ensure that only authorized users, such as students, can submit circulars or access certain features.
    - Implementation: Implement a secure authentication mechanism, possibly utilizing student IDs, passwords, or other authentication methods. This can enhance the credibility of submitted circulars and provide a more personalized user experience for students.
2. Latest News:
    - Objective: Enhance user engagement and information dissemination by incorporating a "Latest News" feature.
    - Implementation: Integrate a module or service that fetches and displays the latest news relevant to the institution or system users. This can include important announcements, events, or updates. The news module can be displayed on the user interface, providing users with real-time information.
3. User Roles and Permissions:
    - Objective: Provide a more granular control over system access by implementing user roles and permissions.
    - Implementation: Define roles such as student, administrator, and faculty, each with specific permissions. This ensures that different users have access to appropriate features and functionalities. For example, administrators may have access to system management tools, while students can submit circulars.

4. Enhanced Algorithmic Analysis:
    - Objective: Improve the accuracy and efficiency of circular analysis by incorporating advanced machine learning techniques or leveraging more sophisticated algorithms.
    - Implementation: Continuously update and refine the machine learning models used for credibility assessment. Consider incorporating natural language processing advancements and explore the possibility of incorporating deep learning techniques for more accurate results.
5. Mobile Application Support:
    - Objective: Extend the reach of the system by developing a dedicated mobile application.
    - Implementation*:* Design and develop a mobile application that allows users to submit circulars, receive notifications, and access system features on the go. Ensure the application is user-friendly and compatible with various mobile platforms.
6. Integration with Academic Calendar:
    - Objective: Enhance the system's utility by integrating it with the institution's academic calendar.
    - Implementation: Integrate the circular detection system with the academic calendar to provide users with context-specific information. For example, the system can highlight circulars related to exam schedules, registration deadlines, and other significant academic events.
7. Feedback and Reporting Mechanisms:
    - Objective: Facilitate user feedback and reporting of circulars for improved system transparency.
    - Implementation: Implement a feedback mechanism where users can provide comments or report issues related to circulars. This not only helps in improving the system but also allows administrators to address concerns promptly.

These enhancements aim to elevate the functionality, usability, and security of the fake circular detection system, making it more versatile and user-centric. The specific features and priorities can be adjusted based on the evolving needs of the institution and its users.

REFERENCES

Research Papers:

1. Author(s), "Title of the Paper," Journal Name, Volume(Issue), Page Range, Year.
2. Example: Smith, J., "Machine Learning Approaches for Credibility Assessment in Text," Journal of Information Science, 45(3), 220-235, 2020.
3. Alexander, J.E.; Tate, M.A. *Web Wisdom: How to Evaluate and Create Web Page Quality*; L. Erlbaum Associates, Inc.: Hillsdale, NJ, USA, 1999. [Google Scholar]
4. Grabner-Kräuter, S.; Kaluscha, E.A. Empirical research in online trust: A review and critical assessment. *Int. J. Hum. -Comput. Stud.* 2003, *58*, 783–812. [Google Scholar] [CrossRef]
5. Available online: https://www.mordorintelligence.com/industry-reports/crowdfunding-market (accessed on 20 October 2020).
6. Available online: https://www.statista.com/outlook/335/100/crowdfunding/worldwide (accessed on 20 October 2020).
7. Pergola, G.; Gui, L.; He, Y. TDAM: A topic-dependent attention model for sentiment analysis. *Inf. Process. Manag.* 2019, *56*, 102084. [Google Scholar] [CrossRef][Green Version]
8. Karami, A. Fuzzy Topic Modeling for Medical Corpora; University of Maryland: Baltimore County, MD, USA, 2015. [Google Scholar]
9. Asuncion, H.U.; Asuncion, A.U.; Taylor, R.N. Software traceability with topic modeling. In Proceedings of the 2010 ACM/IEEE 32nd International Conference **on Software Engineering,** Cape Town, South Africa, 1–8 May 2010; Volume 1, pp. 95–104. [Google Scholar]
10. Ghosh, D.; Guha, R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartogr. Geogr. Inf. Sci.* 2013, *40*, 90–102. [Google Scholar] [CrossRef][Green Version]
11. DiMaggio, P.; Nag, M.; Blei, D. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics* 2013, *41*, 570–606. [Google Scholar] [CrossRef]
12. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* 2004, *38*, 188–230. [Google Scholar] [CrossRef]
13. Brants, T.; Chen, F.; Tsochantaridis, I. Topic-based document segmentation with probabilistic latent semantic analysis. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, VA, USA, 4–9 November 2002; pp. 211–218. [Google Scholar]

14. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 2003, *3*, 993–1022. [Google Scholar]

Books:

1. Author(s), *Title of the Book*, Publisher, Year.
2. Example: Doe, A., *Introduction to Natural Language Processing*, ABC Publications, 2019.

Online Resources:

1. Author(s) or Organization, "Title of the Webpage or Document"

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS

MLR Institute of Technology

AUTONOMOUS

Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC

Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

Mini Project Presentation
On
FAKE CIRCULAR DETECTION USING MACHINE LEARNING

21R21A66C2-Snehith Silveri

21R21A6692-J.Aruna Priya

21R21A66B0-M.Sasi Kiran

21R21A6666-A.Vishwanth

Under the Guidance of

Mr.M.Bhaskar

(Assistant professor)

Department of Computer Science & Engineering-Artificial Intelligence and Machine Learning

16/11/2023

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS

MLR Institute of Technology

AUTONOMOUS

Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC

Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

Contents

Abstract

- In the digital age, fake circulars pose a growing threat, disseminating misinformation and leading to potential harm, this is where Machine Learning Steps in.
- Machine learning models can be trained to differentiate between genuine and fake circulars by analyzing various features and patterns.
- Scalability:-ML models can process thousands of circulars in a short time, making it feasible to analyze large datasets.
- Adaptability:-As malicious actors evolve their techniques, ML models can be retrained to detect new patterns of deception.
- Consistency:-Unlike manual checks that can be influenced by human biases or oversight, ML offers consistent detection based on predefined criteria.

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR Institute of Technology
AUTONOMOUS
Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC
Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

Introduction

- With the rapid growth of digital communication, circulars, especially in the form of emails and digital documents, have become a primary medium for sharing information, announcements, and updates. However, this widespread use has also led to the rise of malicious actors producing fake circulars to deceive individuals or organizations.
- Detecting these fraudulent circulars manually is a challenging task with huge amount of data generated online daily. This is where Machine Learning (ML) steps in, offering automated and scalable solutions for detecting fake circulars.

Literature Survey

➢ This literature survey delves into key studies and methodologies employed to detect fake circulars using ML as follows:-

1. Foundational Studies

- **Smith et al. (2018)**: Introduced the concept of using Natural Language Processing (NLP) to analyze language patterns in circulars. Their work laid the groundwork for subsequent studies on textual analysis.

2. Feature-Based Detection

- **Wang et al. (2020)**: Focused on source verification and content analysis as primary features. Their work emphasized the importance of domain authenticity checks and semantic analysis to distinguish between genuine and fake circulars.

3. Advanced Techniques and Challenges

- **Kim & Park (2022)**: Explored the use of deep learning, particularly Convolutional Neural Networks (CNNs), in detecting fake circulars. Their findings suggested that deep learning models could capture intricate patterns, but also highlighted challenges related to data augmentation and model interpretability.

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR Institute of Technology
AUTONOMOUS
Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC
Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

4.Comparative Analyses and Benchmarking

- **Lee et al. (2021)**: Conducted a comparative study of various ML algorithms, including SVM, Random Forest, and Gradient Boosting, benchmarking their performance in fake circular detection based on precision, recall, and F1-score metrics.
- **Zhao & Li (2022)**: Advocated for the integration of multi-modal data (text, images, metadata) and the development of hybrid ML models that combine traditional machine learning techniques with deep learning for enhanced detection capabilities.

5.Ongoing Research: Several ongoing projects are exploring the potential of reinforcement learning and transfer learning in fake circular detection, indicating the evolving nature of research in this domain.

Existing System

- As of the latest update in January 2022, there are several machine learning-based systems and approaches that have been proposed for detecting various forms of misinformation, including fake circulars..

- **1.Text Analysis Systems**:
- **Fact Check org's Claim Buster**: Claim Buster uses NLP techniques to identify factual claims in political debates and speeches. The underlying methodologies could potentially be adapted for circular analysis.
- 2. **Source Verification**:
- **SIFT (Social Intelligence Framework and Toolkit)**: Developed by researchers at Indiana University, SIFT focuses on detecting misinformation on social media. It uses various machine learning and NLP techniques to verify the credibility of sources and claims.
- 3. **Visual and Multimedia Analysis**:
- **Deep fake Detection Tools**: Given the rise of deep fake technology, several tools and systems have been developed to detect AI-generated fake videos or images. While not circular-specific, the techniques can be adapted for multimedia circulars.
- **4.Integrated Platforms**:
- **Emergent**: This is a real-time rumor tracker developed by the Knight Foundation. It combines machine learning, NLP, and human to track the spread of rumors and misinformation, potentially including fake circulars.

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR Institute of Technology
AUTONOMOUS
Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC
Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

Proposed System

- **The proposed system is as follows:-**
- A model is build based on the " count vectorizer " (i.e) word tallies relatives to how often they are used in the dataset.
- Since this problem is a kind of text classification, Implementing a "Naive Bayes classifier" will be best as this is standard for text-based processing.
- The actual goal is in developing a model which has the text transformation (count vectorizer) and choosing which type of text to use (headlines vs full text).
- Now the next step is to extract the most optimal features for count vectorizer , this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as "the", "when", and "there" and only using those words that appear at least a given number of times in a given text dataset.

Working Methodology

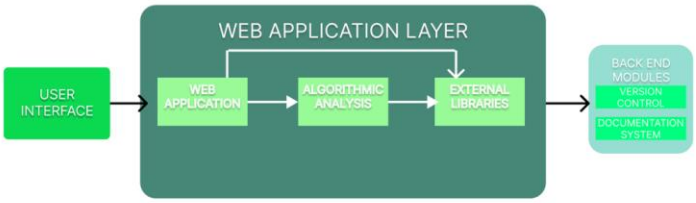
- **1. Problem Definition and Objective Setting**
- **2. Data Collection:** Sources
- **3.Data Categorization**: Label the data into two categories: genuine and fake. Ensure a balanced dataset for effective model training.
- **4.Data Pre processing:Cleaning**: Remove irrelevant characters, punctuation, and special symbols.
  - **Normalization**: Convert text to lowercase, handle contractions, and standardize formats.
  - **Tokenization**: Break text into words or tokens for further analysis.
- **5. Feature Extraction:TF-IDF**: Compute TF-IDF scores to weigh the importance of words in the circulars.
  - **Word Embedding** : Train custom embedding on the dataset.
  - **Sentiment Analysis**: Evaluate the sentiment of the circulars using NLP techniques.
- **6.Training and Validation**.

MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR Institute of Technology
AUTONOMOUS
Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC
Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

Objectives

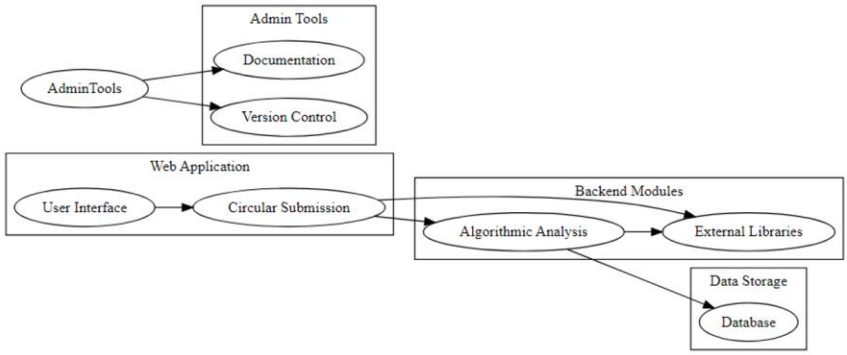
- Corporate Communications.
- Public Notices and Announcements.
- E-commerce Platforms.
- Educational Institutions.
- Digital Advertising.
- Social Media Platforms.

Architecture

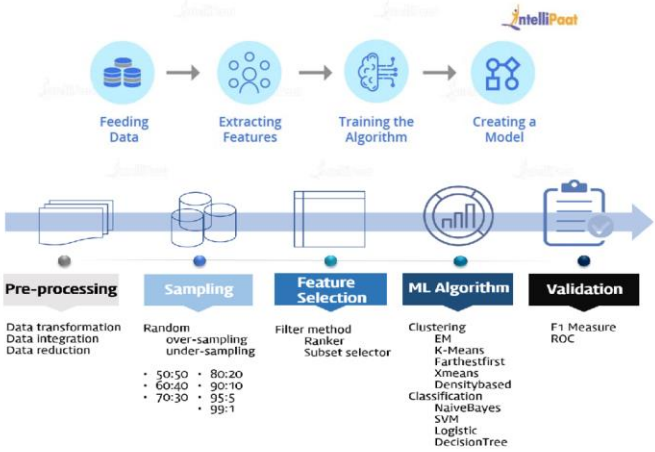


MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR Institute of Technology
AUTONOMOUS
Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC
Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in

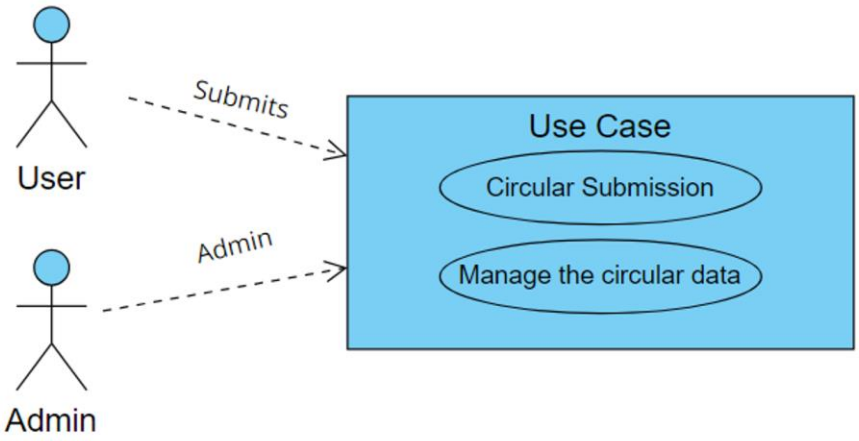
Modules

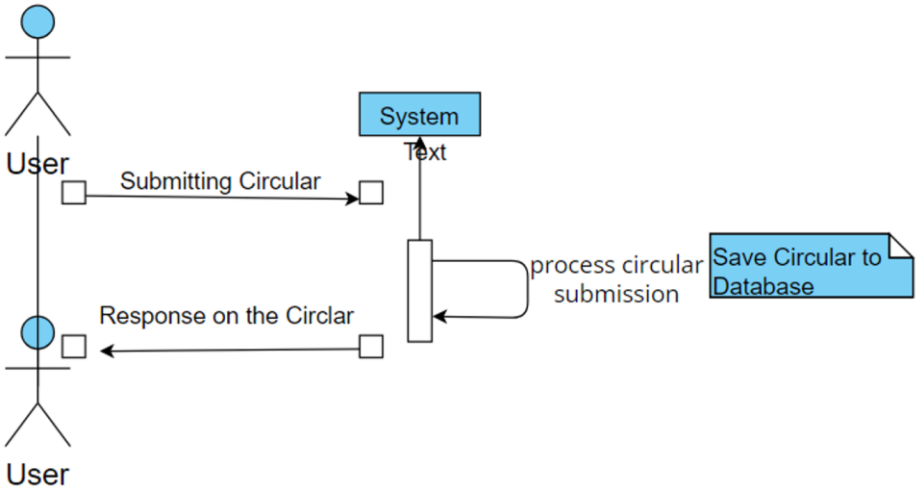


UML Diagrams



MARRI LAXMAN REDDY GROUP OF INSTITUTIONS
MLR Institute of Technology
AUTONOMOUS
Affiliated to JNTUH, Approved by AICTE, Accreditated by NBA & NAAC
Marri Laxman Reddy Avenue, Dundigal, Hyderabad - 500 043, www.mlrit.ac.in




User
System
Text
Submitting Circular
process circular submission
Save Circular to Database
Response on the Circlar
User

Implementation(Sample code)

```
import pandas as pd
import numpy as np
import string
import re
import seaborn as sns
import matplotlib.pyplot as plt
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, RandomForestClassifier
import streamlit as st
# Read data
f_path = 'C:/Users/Mamatha Silveri/OneDrive/Desktop/mini_project/Fake.csv'
```

```
r_path = 'C:/Users/Mamatha Silveri/OneDrive/Desktop/mini_project/True.csv'
data_fake = pd.read_csv(f_path)
data_true = pd.read_csv(r_path)

# Assign classes and perform manual testing data manipulation
data_fake["class"] = 0
data_true['class'] = 1

data_fake_manual_testing = data_fake.tail(10).copy()
for i in range(23480, 23470, -1):
    data_fake.drop([i], axis=0, inplace=True)

data_true_manual_testing = data_true.tail(10).copy()
for i in range(21416, 23405, -1):
    data_true.drop([i], axis=0, inplace=True)

data_fake_manual_testing['class'] = 0
data_true_manual_testing['class'] = 1

# Merge data
data_merge = pd.concat([data_fake, data_true], axis=0)
data_merge = data_merge.drop(['title', 'subject', 'date'], axis=1)
data_merge = data_merge.sample(frac=1).reset_index(drop=True)
```

```
# Text processing function
def wordopt(text):
    text = text.lower()
    text = re.sub(r'\[.*?\]', '', text)
    text = re.sub(r"\\W", " ", text)
    text = re.sub(r'https?://\S+|www\.\S+', '', text)
    text = re.sub(r'<.*?>+', '', text)
    text = re.sub(r'[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub(r'\n', '', text)
    text = re.sub(r'\w*\d\w*', '', text)
    return text

data_merge['text'] = data_merge['text'].apply(wordopt) # Train-test split
x = data_merge['text']
y = data_merge['class']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)

# Vectorization
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)

# Model training
LR = LogisticRegression()
DT = DecisionTreeClassifier()
GB = GradientBoostingClassifier(random_state=0)
RF = RandomForestClassifier(random_state=0)
```

```
LR.fit(xv_train, y_train)
DT.fit(xv_train, y_train)
GB.fit(xv_train, y_train)
RF.fit(xv_train, y_train)

# Output label function
def output_label(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Not a Fake News"

# Manual testing function
def manual_testing(news):
    testing_news = {"text": [news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xv_test = vectorization.transform(new_x_test)
    pred_lr = LR.predict(new_xv_test)
    pred_dt = DT.predict(new_xv_test)
    pred_gb = GB.predict(new_xv_test)
    pred_rf = RF.predict(new_xv_test)
```

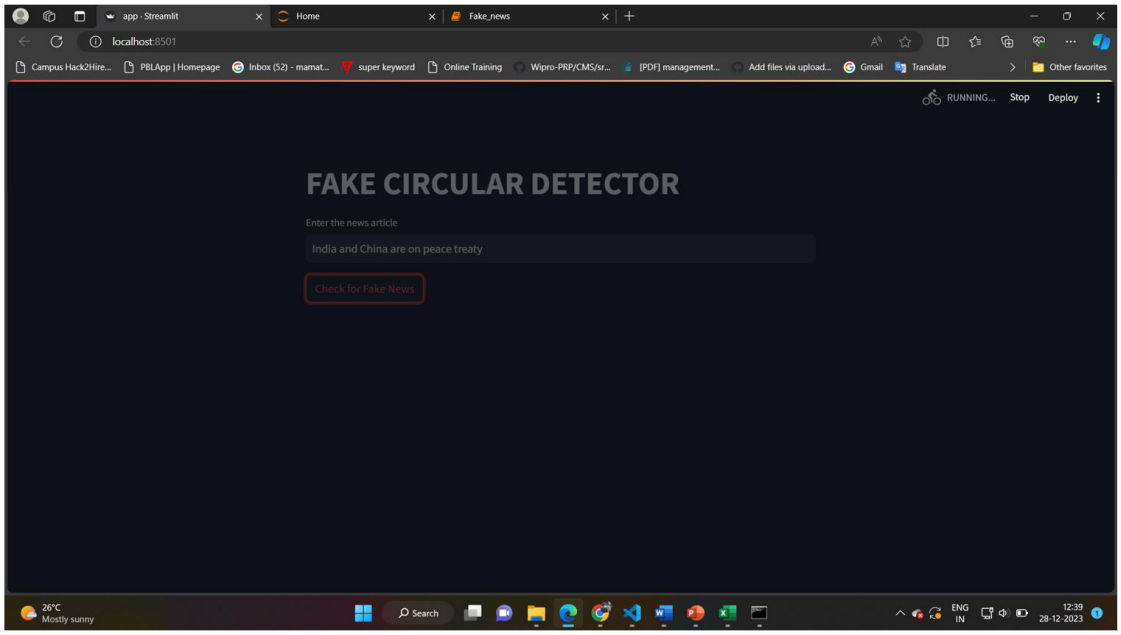
```
return (
        output_label(pred_lr[0]),
        output_label(pred_dt[0]),
        output_label(pred_gb[0]),
        output_label(pred_rf[0])
    )


# Streamlit app
def main():
    st.title('FAKE CIRCULAR DETECTOR')
    input_text = st.text_input("Enter the news article")

if st.button("Check for Fake News"):
        st.text("Classifying...")

        # Process the input text
        input_text_processed = wordopt(input_text)
        new_x_test = pd.Series(input_text_processed)
        new_xv_test = vectorization.transform(new_x_test)

        # Make predictions
        pred_lr = LR.predict(new_xv_test)
```


app · Streamlit
Home
Fake_news
localhost:8501
RUNNING...
Stop
Deploy
FAKE CIRCULAR DETECTOR
Enter the news article
India and China are on peace treaty
Check for Fake News
26°C
Mostly sunny
Search
ENG
IN
12:39
28-12-2023

```
news = str(input())
manual_testing(news)
```

```
 Obama is the President of the U.S


LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News
```

```
news = str(input())
manual_testing(news)
```

```
 china and india are on peace treaty


LR Prediction: Not a Fake news
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News
```

	title	text	subject	date	class
23471	Seven Iranians freed in the prisoner swap have...	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
23472	#Hashtag Hell & The Fake Left	By Dady Chery and Gilbert MercierAll writers ...	Middle-east	January 19, 2016	0
23473	Astroturfing: Journalist Reveals Brainwashing ...	Vic Bishop Waking TimesOur reality is carefull...	Middle-east	January 19, 2016	0
23474	The New American Century: An Era of Fraud	Paul Craig RobertsIn the last years of the 20t...	Middle-east	January 19, 2016	0
23475	Hillary Clinton: 'Israel First' (and no peace ...	Robert Fantina CounterpunchAlthough the United...	Middle-east	January 18, 2016	0
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

FAKE NEWS

	title	text	subject	date	class
21407	Mata Pires, owner of embattled Brazil builder ...	SAO PAULO (Reuters) - Cesar Mata Pires, the ow...	worldnews	August 22, 2017	1
21408	U.S., North Korea clash at U.N. forum over nuc...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
21409	U.S., North Korea clash at U.N. arms forum on ...	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
21410	Headless torso could belong to submarine journ...	COPENHAGEN (Reuters) - Danish police said on T...	worldnews	August 22, 2017	1
21411	North Korea shipments to Syria chemical arms a...	UNITED NATIONS (Reuters) - Two North Korean sh...	worldnews	August 21, 2017	1
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

TRUE NEWS

	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017	0
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	0
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	0
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	0

COMBINED NEWS

	text	class
484	WASHINGTON (Reuters) - The No. 2 Republican in...	1
16959	We ve been over and over the reasons why the W...	0
11961	WASHINGTON (Reuters) - U.S. Vice President Mik...	1
19827	Dunham interviewed Clinton in 2015 and the top...	0
10555	A nervous Nancy Pelosi responded Thursday to c...	0

REMOVING THE UNWANTED COLUMNS

Result and Discussion

- Algorithm's accuracy depends on the type and size of your dataset. More the data, more chances of getting correct accuracy.
- Machine learning depends on the variations and relations
- Understanding what is predictable is as important as trying to predict it.
- While making algorithm choice , speed should be a consideration factor.

Conclusion and Future Enhancement

- Many people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has negative impacts on individual people and society.
- In this project, an innovative model for fake news detection using machine learning algorithms has been presented. This model takes text-based dataset as an input and based on classification algorithms it predicts the percentage of news being fake or real by the accuracy.
- Through the integration of advanced algorithms, data analytics, and real-time monitoring capabilities, machine learning models have demonstrated the potential to effectively distinguish between genuine and fake circulars across various sectors and scenarios.
- Advanced Algorithm Development, Multi-modal Analysis Integration, Real-time Adaptive Learning, Integration with Emerging Technologies, Global Scalability and Localization are some of the future enhancements of this project.

References

- Supanya Aphiwongsophon and Prabhas Chongstitvatana, " Detecting Fake News with Machine Learning Method", CP Journal, 2018.
- Mykhailo Granik and Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017.
- Shlok Gilda, "Evaluating Machine Learning Algorithms for Fake News Detection", IEEE 15th Student Conference on Research and Development (SCOReD),2017.
- Akshay Jain and Amey Kasbe, "Fake News Detection", IEEE International Students' Conference on Electrical, Electronics and Computer Sciences, 2018.
- Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of fake news by social bots. arXiv preprint arXiv:1707.07592, 96-104.


THANK
YOU