A

Project Report

On

FAKE NEWS DETECTION WITH MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of Degree

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING (AI&ML)

by

JIYA GARG	207R1A6688
ABHINAV LAKKARAJU	207R1A6698
SANJAY KATUKOJWALA	207R1A6695

Under the Guidance of

MR. SK. SHARIF

(Assistant professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (AI&ML)

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi) Recognized Under Section 2(f) & 12(B) of the UGCAct.1956, Kandlakoya (V), Medchal Road, Hyderabad-501401.

2020-2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (AI&ML)

CERTIFICATE

This is to certify that the project entitled **"FAKE NEWS DETECTION WITH MACHINE LEARNING"** being submitted by **JIYA GARG (207R1A6688), ABHINAV LAKKARAJU(207R1A6698) & SANJAY KATUKOJWALA (207R1A6695)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering (AI&ML) to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by them under our guidance and supervision during the year 2023-24.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Mr.SK. Sharif
Assistant Professor
INTERNAL GUIDE

Dr. S Rao Chintalapudi
HOD CSE(AI&ML)

EXTERNAL EXAMINER

Submitted for viva voice Examination held on _____

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

We take this opportunity to express my profound gratitude and deep regard to my guide **Mr. Sk. Sharif,** Associate Professor and HOD CSE(AI&ML) for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to the Project Review Committee (PRC) **Dr. G. Vinoda Reddy, Dr. K. Mahesh, N. Sateesh & B. Mamatha** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. S Rao Chintalapudi,** Head, Department of Computer Science and Engineering (AI&ML) for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We also express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

JIYA GARG (207R1A6688)

ABHINAV LAKKARAJU (207R1A6698)

SANJAY KATUKOJWALA (207R1A6695)

ABSTRACT

The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it.

The rise of ubiquitous deepfakes, misinformation, disinformation, and post-truth, often referred to as fake news, raises concerns over the role of the Internet and social media in modern democratic societies. Due to its rapid and widespread diffusion, digital deception has not only an individual or societal cost, but it can lead to significant economic losses or to risks to national security.

Blockchain and other distributed ledger technologies (DLTs) guarantee the provenance and traceability of data by providing a transparent, immutable, and verifiable record of transactions while creating a peer-to-peer secure platform for storing and exchanging information.

This over view aims to explore the potential of DLTs to combat digital deception, describing the most relevant applications and identifying their main open challenges. Moreover, some recommendations are enumerated to guide future researchers on issues that will have to be tackled to strengthen the resilience against cyber-threats on today's online media.

This project makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python, NLP for textual analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.

LIST OF FIGURES

TABLE NO	TABLE NAME	PAGE NO
3.1	Project Architecture of Fake News Detection with Machine Learning	10
3.2	Use Case Diagram for Fake News Detection with Machine Learning	12
3.3	Class Diagram for Fake News Detection Using Machine Learning	14
3.4	Sequence Diagram Fake News Detection Using Machine Learning	15
4.4	Dataset Description	21
4.7	Results Analysis	26
5.2	User Login Page	28
5.3	Upload the dataset	29
5.4	Run The Module	29
5.5	Result with rank score and Accuracy	30

LIST OF TABLES

TABLE NO	TABLE NAME	PAGE NO
6.3	TESTCASES	33

TABLE OF CONTENTS

ABSTRACT i

LIST OF FIGURES ii

LIST OF TABLES iii

1.INTRODUCTION **1**

1.1 PROJECT SCOPE 1

1.2 PROJECT PURPOSE 2

1.3 PROJECT FEATURES 2

2.SYSTEM ANALYSIS **4**

2.1 PROBLEM DEFINITION 5

2.2 EXISTING SYSTEM / LITERATURE REVIEW 6

2.2.1 EXISTING SYSTEM-1 6

2.2.2 EXISTING SYSTEM-2 6

2.2.3 LIMITATIONS OF EXISTING SYSTEMS 6

2.3 PROPOSED SYSTEM 7

2.3.1 PROPOSED APPROACH 7

2.3.2 ADVANTAGES OF PROPOSED SYSTEM 7

2.4 HARDWARE & SOFTWARE REQUIREMENTS 8

2.4.1 HARDWARE REQUIREMENTS 8

2.4.2 SOFTWARE REQUIREMENTS 8

3.ARCHITECTURE **9**

3.1 PROJECT ARCHITECTURE 10

3.2 USE CASE DIAGRAM 12

3.3 CLASS DIAGRAM 17

3.4 SEQUENCE DIAGRAM 15

4.IMPLEMENTATION **17**

4.1 NAIVE BAYES ALGORITHM 18

4.2 ANOMALY DETECTOR ALGORITHM 19

4.3 NATURAL LANGUAGE PROCESSING 20

4.4 DATASET DESCRIPTION 21

4.5 PERFORMANCE METRICS 22

4.6 SAMPLE CODE 23

4.7 RESULT ANALYSIS 26

5.SCREENSHOTS **27**

6.TESTING **31**

6.1 INTRODUCTION TO TESTING 32

6.2 TYPES OF TESTING 32

6.2.1 UNIT TESTING 32

6.2.2 INTEGRATION TESTING 32

6.2.3 FUNCTIONAL TESTING 33

6.3 TEST CASES 33

7.CONCLUSION & FUTURE SCOPE **34**

7.1 CONCLUSION 35

7.2 FUTURE SCOPE 35

8.BIBLIOGRAPHY **36**

8.1 REFERENCES 37

8.2 GITHUB LINK 37

1. INTRODUCTION

1. INTRODUCTION

1.1 PROJECT SCOPE

The scope of this project involves leveraging Natural Language Processing (NLP) techniques to classify fake news articles by focusing on identifying in-article attribution. In-article attribution refers to the practice of attributing information or quotes to specific sources within the text of an article. By analyzing these attributions, the project aims to develop a supervised learning estimator that can differentiate between genuine and fake news articles.

To achieve this, the project will involve collecting a dataset of news articles labeled as either genuine or fake. The dataset will be preprocessed to extract relevant features, with a particular emphasis on identifying and analyzing in-article attributions. Various NLP techniques, such as named entity recognition and sentiment analysis, will be applied to extract and interpret information from the text. Machine learning algorithms will then be trained on these features to build a classification model capable of distinguishing between real and fake news articles based on their attribution patterns. Various NLP techniques, such as named entity recognition and sentiment analysis, will be applied to extract and interpret information from the text. . Machine learning algorithms will then be trained on these features to build a classification model capable of distinguishing between real and fake news articles based on their attribution patterns

The scope also encompasses evaluating the performance of the developed model using appropriate metrics and potentially refining the model based on feedback and further analysis. Additionally, considerations will be made for scalability and generalizability of the model to different types of news articles and sources.

1.2 PROJECT PURPOSE

The purpose of this project is to leverage Natural Language Processing (NLP) techniques to classify fake news articles by focusing on the identification of in-article attribution. In today's information age, the spread of misinformation poses a significant challenge, leading to potentially harmful consequences such as the manipulation of public opinion, erosion of trust in media, and even societal unrest. In-article attribution, the practice of correctly attributing information and sources within articles, is a crucial aspect in discerning the credibility and reliability of news content. By developing a supervised learning estimator, we aim to create a tool capable of automatically analyzing news articles and determining their authenticity based on the presence and quality of in-article attribution.

The primary goal is to contribute to the ongoing efforts in combating the proliferation of fake news. By focusing on in-article attribution, we aim to develop a more nuanced approach to identifying misinformation, which goes beyond simple keyword analysis or sentiment classification.

By leveraging NLP techniques and supervised learning, the project seeks to provide a scalable solution that can assist journalists, fact-checkers, and the general public in identifying fake news articles more efficiently. By accurately flagging suspicious articles, the tool can help mitigate the harmful effects of misinformation on public discourse, political processes, and societal well-being.

This project offers an opportunity to explore and advance NLP techniques, particularly in the domain of text classification and information extraction. By tackling the challenge of identifying in-article attribution, we aim to contribute to the broader field of NLP research and develop methodologies that could be applicable beyond the scope of fake news detection.

Furthermore, by shedding light on the role of in-article attribution in distinguishing between genuine and fake news, this project contributes to a deeper understanding of the linguistic and semantic features of deceptive content, thereby advancing research in the field of computational journalism and NLP.

.

1.3 PROJECT FEATURES

The endeavor to classify fake news articles using Natural Language Processing (NLP) with a focus on identifying in-article attribution entails a multifaceted approach, integrating various features and functionalities to achieve its objectives effectively. Below are the key features of this project:

the project lies the development of a supervised learning model. This model will be trained on labeled datasets comprising both authentic and fake news articles. Through the utilization of algorithms Neural Networks, the model will learn to classify articles based on the presence and quality of in-article attribution. The training process involves the extraction of relevant features from the text data and mapping them to corresponding labels, enabling the model to generalize patterns and make accurate predictions on unseen articles.

Leveraging a spectrum of NLP techniques is crucial for extracting meaningful information from text data. Tokenization breaks down articles into individual words or phrases, enabling further analysis. Named Entity Recognition (NER) identifies entities such as names of people, organizations, and locations, which can aid in assessing the credibility of sources mentioned in articles. Part-of-Speech (POS) tagging assigns grammatical categories to words, facilitating syntactic analysis. Additionally, sentiment analysis discerns the emotional tone of the text, which may provide insights into the subjective nature of the content.

the project is committed to continuous improvement and adaptation. Feedback mechanisms and user engagement initiatives facilitate iterative enhancements to the classification model and the overall system. Regular updates address emerging challenges and evolving trends, ensuring the relevance and effectiveness of the classification tool over time.

Incorporating these features enables the project to develop a robust and versatile solution for classifying fake news articles, contributing to the broader efforts in combating misinformation and promoting media literacy.

2.SYSTEM ANALYSIS

2 SYSTEM ANALYSIS

System Analysis is the important phase in the system development process. The System is studied to the minute details and analyzed. The system analyst plays an important role of an interrogator and dwells deep into the working of the present system. In analysis, a detailed study of these operations performed by the system and their relationships within and outside the system is done. A key question considered here is, “what must be done to solve the problem?” The system is viewed as a whole and the inputs to the system are identified. Once analysis is completed the analyst hasa firm understanding of what is to be done.

2.1 PROBLEM DEFINITION

Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator" revolves around developing a system that can effectively differentiate between authentic news articles and fake ones by focusing on in-article attribution, using supervised learning techniques in Natural Language Processing (NLP). a classification system capable of distinguishing between authentic news articles and fake ones. This entails training a supervised learning model using NLP techniques on a labeled dataset of news articles, where each article is categorized as authentic or fake based on its content Natural Language Processing (NLP) techniques are used to extract relevant features from the text of each news article. In this project, there is a specific focus on in-article attribution, which refers to the identification of the sources or references cited within the article. Other features such as word frequencies, sentiment analysis, and syntactic structures may also be considered.

2.2 EXISTING SYSTEM 1

Fake news has been demonstrated to be problematic in multiple ways. It has been shown to have real influence on public perception and the ability to shape regional and national dialogue. It has harmed business and individuals and even resulted in death, when an individual responded to a hoax. It has caused some teenagers to reject the concept of media objectivity and many students can't reliably tell the difference between real and faked articles. It is even thought to have influenced the 2016 United States elections. Fake news can be spread deliberately by humans or indiscriminately by bot armies, with the latter giving a nefarious article significant reach. Not just articles are faked, in many cases fake, mislabeled or deceptive images are also used to maximize impact. Some contend that fake news is a "plague" on society's digital

infrastructure. Many are working to combat it. Farajtabar, et al., for example, has proposed a system based on points, while Haigh, Haigh and Kozakhave suggested the use of "peer-to-peer counter propaganda

2.2.2 EXISTING SYSTEM 2

- The existing system for combating fake news primarily relies on human Effort fact-checkers and journalists.
- This process can be time-consuming and labor-intensive in case of large data.
- Different organizations use different criteria for determining news is true or false,this can lead to inaccuracies.

2.2.3 LIMITATIONS OF EXISTING SYSTEMS

Following are the disadvantages of existing system:

- Absence of learning algorithms like SVM and Naive Bayes.
- Failure to address social networking issues such as privacy concerns, online bullying, misuse, and trolling.

2.3 PROPOSED SYSTEM

2.3.1 Proposed Approach

In This project author is describing concept to detect fake news from social media or document corpus using Natural Language Processing and attribution supervised learning estimator. News documents or articles will be uploaded to application and then by using Natural Language Processing to extract quotes, verbs and name entity recognition (extracting organizations or person names) from documents to compute score, verbs, quotes and name entity also called as attribution. Using supervised learning estimator, we will calculate score between sum of verbs, sum of name entity and sum of quotes divided by total sentence length. If score greater than 0.9 then news will be considered as REAL and if less than 0.9 then new will be consider as FAKE.

2.3.2 ADVANTAGES OF THE PROPOSED SYSTEM

The proposed system implemented using the machine learning techniques, the proposed system is processing in the following way.

- Improved accuracy in fake News identification
- Comprehensive analysis considering various factors
- Adaptive learning capability for staying effective over time
- Increased efficiency through automation
- Minimization of false positives
- Effective handling of diverse textual data with NLP

2.4 HARDWARE & SOFTWARE REQUIREMENTS

2.4.1 HARDWARE REQUIREMENTS:

Hardware interfaces specify the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

- PROCESSOR : i5 or above
- RAM : 8GB (min)
- HARD DISK : 256 GB
- KEYBOARD : Standard Windows Keyboard
- MOUSE : Two or Three Button Mouse
- MONITOR : HDMI

2.4.2 SOFTWARE REQUIREMENTS:

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements,

- OPERATING SYSTEM : Windows 10
- CODE LANGUAGE : Python
- LIBRARIES : Django libraries
  Re, Textblob and Nltk
- FRONT-END : Python

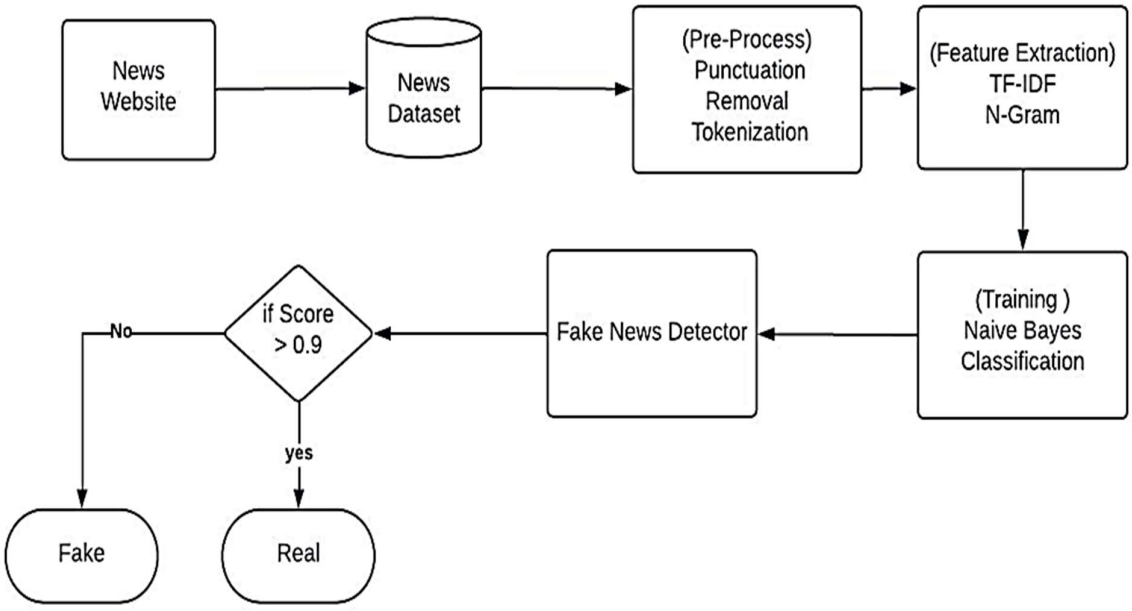
- BACK-END : Django-ORM
- DESIGNING : HTML, CSS
- Web Server : Django Development Server

3. ARCHITECTURE

3.ARCHITECTURE

3.1 PROJECT ARCHITECTURE

This project architecture shows the procedure followed for classification, starting from input to final prediction



3.1: Project Architecture of Fake News Detection with Machine Learning

DESCRIPTION

- **Data Collection**: In this stage, news articles are collected from various sources, including news websites and social media.
- **Pre-processing**: The collected data undergoes pre-processing to get it ready for analysis by the machine learning model. This may involve removing punctuation, converting text to lowercase, and tokenization, which is breaking the text down into individual words.
- **Feature Extraction**: Features are extracted from the pre-processed text data. These features are characteristics that the machine learning model will use to identify patterns that differentiate real news from fake news. Some common features used for fake news detection include:
    - **TF-IDF**: Term frequency-inverse document frequency is a statistic that reflects how important a word is to a document in a collection.
    - **N-grams**: These are sequences of N words that can be helpful in capturing the phrasing and stylistic choices used in the text.
- **Training**: The extracted features are used to train a machine learning model, such as a Naive Bayes classifier. During training, the model is provided with labeled data sets of real and fake news articles. The model learns to identify the characteristics that differentiate real from fake news based on these examples.
- **Fake News Detection**: Once the model is trained, it can be used to detect fake news in new articles. The features are extracted from the new article, and the model is used to classify the article as real news or fake news.
- **Thresholding**: The system might have a threshold in place to determine how likely an article is to be fake news. For instance, if the score is greater than 0.9, the news article is classified as fake news.

3.2 USE CASE DIAGRAM

In the use case diagram, we have basically one actor who is the user in the trained model.

A use case diagram is a graphical depiction of a user's possible interactions with a system. A use case diagram shows various use cases and different types of usersthe system has. The use cases are represented by either circles or ellipses. The actors are often shown as stick figures.

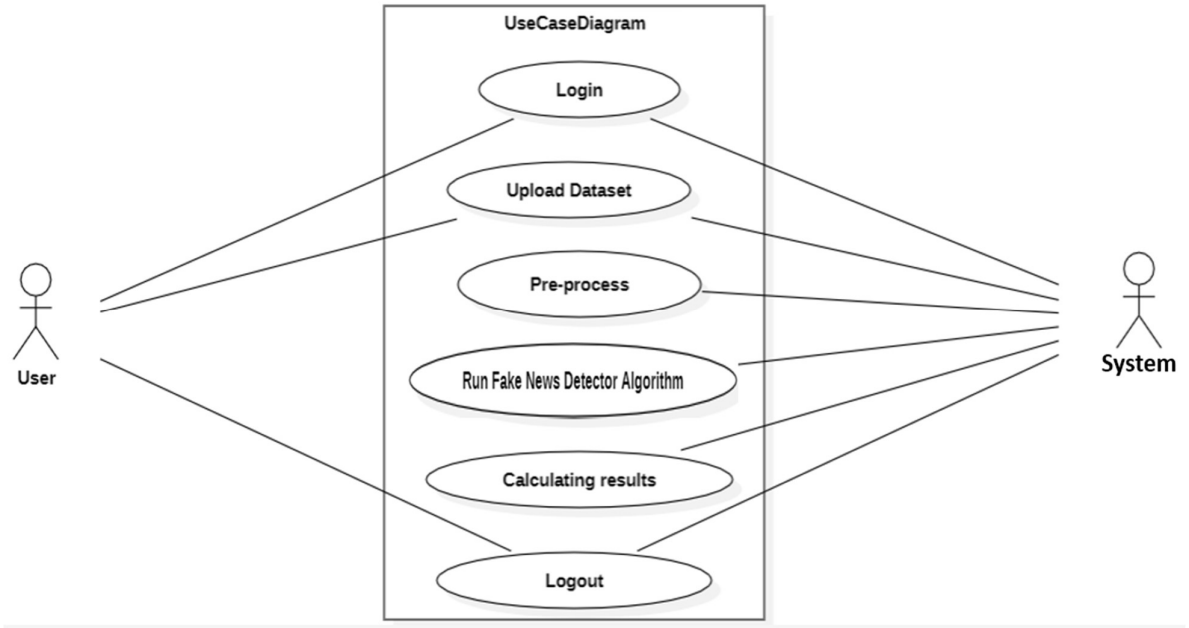


Figure 3.2: Use Case Diagram for Fake News Detection with Machine Learning

DESCRIPTION

A use case diagram in the Unified Modelling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided. by a system in terms of actors, their goals and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actor system can be depicted in the system.

In the diagram our actors are user and system. These actors are performing their own functionalities.

First the user performs the login function. to get user access to the web page and then will upload the news data sets for the purpose to get the desired result. Then the next step is is done by the accessing which preprocessing with the system data in to clean data. Then the system will run the Detector Algorithm which will classify whether the particular news. in fake or real and after getting the result, the user will logout from the home page Thes diagram shows the association relationship b/w the performed by him it.

3.3 CLASS DIAGRAM

Class diagram is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations(or methods), and the relationships among objects.

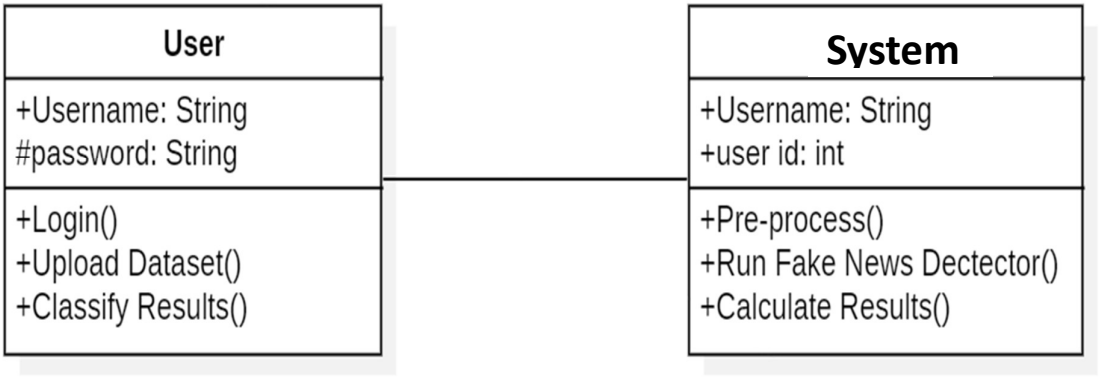


Figure 3.3: Class Diagram for Fake News Detection Using Machine Le

DESCRIPTION

detecting fake news articles using machine learning and natural language processing (NLP). It outlines the interaction between a user and an admin.

- **User:** The user can upload a dataset of news articles and initiate the fake news detection algorithm.
- **Admin:** The admin can pre-process the data, which likely involves cleaning and preparing the text data for analysis. They can also interact with a rule detector algorithm, which might represent the process of training a machine learning model on labeled data (real and fake news articles).

The system might also include a threshold for determining how likely an article is fake news.

.

3.4 SEQUENCE DIAGRAM

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the logical view of the system under development.

SEQUENCE DIAGRAM FOR BULDING FAKE PROFILE IDENTIFICATION MODEL

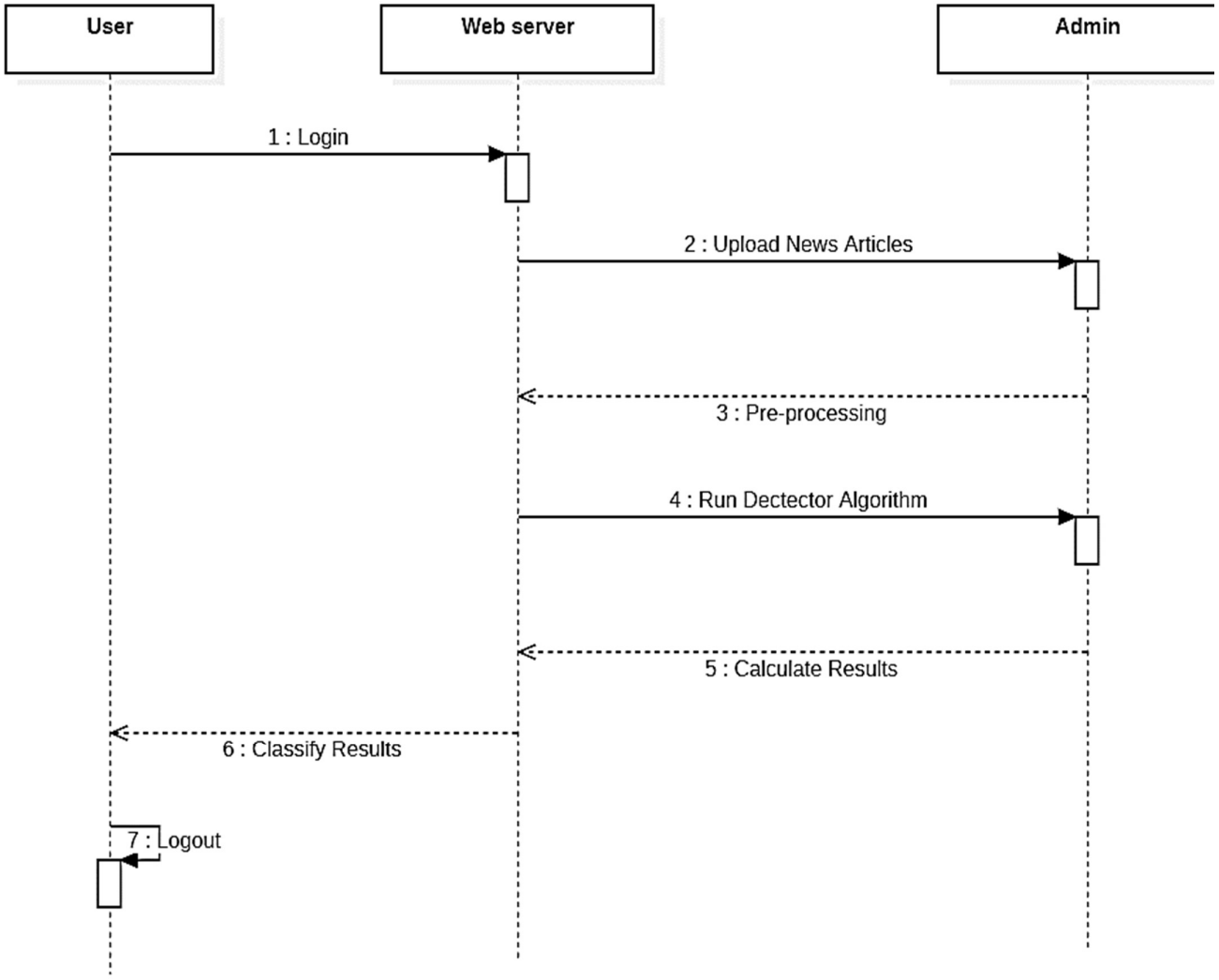


Figure 3.4: Sequence Diagram for building fake news detection model Using Machine Learning

DESCRIPTION

Data Upload: The process starts with uploading news articles, possibly from various sources like social media and news websites.

Pre-processing: The uploaded data undergoes pre-processing steps to get it ready for analysis. This may involve removing punctuation, converting text to lowercase, and breaking the text down into individual words (tokenization).

Feature Extraction: Once the data is cleaned, features are extracted from the text. These features are characteristics that help the machine learning model identify patterns that differentiate real news from fake news. Some examples include:

TF-IDF: This stands for Term Frequency-Inverse Document Frequency. It reflects how important a word is to a specific document in a collection of documents.
N-grams: These are sequences of N words that can help capture the phrasing and stylistic choices used in the text.
Fake News Detection: After feature extraction, a detector algorithm is run to classify the news article. This algorithm is likely a machine learning model that has been trained on a large dataset of labeled real and fake news articles. The model uses the extracted features to classify the new article as real or fake news.

Thresholding (Optional): The system might have a threshold in place to determine how likely an article is fake news. For instance, if the score generated by the model is greater than a certain value, the news article might be classified as fake news.

There are many variations and more sophisticated techniques that can be used for fake news detection. However, it gives you a general idea of the role that machine learning and NLP can play in automating the process of identifying fake news articles.

4. IMPLEMENTATION

4.1 NAIVE BAYES ALGORITHM

Naive Bayes algorithm, a foundational technique in machine learning, operates on the principle of probabilistic inference. It assumes independence among features, meaning that the presence of one feature is considered unrelated to the presence of any other feature, hence the term "naive." This assumption simplifies computation while still offering reasonable performance, especially in text classification tasks like fake news detection

In the context of fake news detection, Naive Bayes calculates the probability that a given piece of news belongs to a particular category (e.g., real or fake) based on the presence of certain words or features in the news article. It estimates the conditional probability of a news article being real (R) or fake (F) given its features (x_1, x_2, ..., x_n) using Bayes' theorem:

$$P(R|X) = \frac{P(X|R) \cdot P(R)}{P(X)}$$

$$P(F|X) = \frac{P(X|F) \cdot P(F)}{P(X)}$$

where:

- $P(R|X)$ is the probability that the news article is real given its features.
- $P(F|X)$ is the probability that the news article is fake given its features.
- $P(X|R)$ is the likelihood of observing the features given that the news article is real.
- $P(X|F)$ is the likelihood of observing the features given that the news article is fake.
- $P(R)$ and $P(F)$ are the prior probabilities of a news article being real or fake, respectively.
- $P(X)$ is the probability of observing the features regardless of the class label.

The class label with the highest posterior probability (either real or fake) is assigned to the news article. Despite its simplicity and the "naive" assumption, Naive Bayes can be surprisingly effective in practice, particularly for text classification tasks, making it a valuable tool in the detection of fake news.

4.2 ANOMALY DETECTOR ALGORITHM

Anomaly detection algorithms are used to identify patterns in data that do not conform to expected behaviour, which may indicate unusual or potentially suspicious activity. One common approach to anomaly detection is based on statistical methods, where anomalies are detected by modelling the normal behaviour of the data and identifying instances that deviate significantly from this model.

One such algorithm is the Gaussian Distribution-based Anomaly Detection Algorithm, also known as the Univariate Anomaly Detection Algorithm. In this algorithm, it is assumed that the features of the data follow a Gaussian (normal) distribution. The algorithm involves two main steps:

1. Model Estimation:

- Given a dataset with m examples and n features, for each feature j, estimate the mean μj and the variance 2σj2.
- These parameters are estimated using the following formulas:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right)$$

2. Anomaly Detection:

- For a new example x, compute the probability p(x) that it belongs to the normal distribution using the multivariate Gaussian probability density function:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right)$$

- Anomaly is detected if p(x) falls below a predefined threshold ε. Instances were p(x)<ε are considered anomalies.
- The threshold ε is typically chosen based on a validation set or cross-validation.

This algorithm is effective for detecting anomalies in datasets where the normal behavior can be modelled well by a Gaussian distribution.

4.3 NATURAL LANGUAGE PROCESSING

In Natural Language Processing (NLP), various algorithms and techniques are employed for tasks such as fake news detection within paragraphs of text. Some commonly used NLP algorithms include:

Word Embeddings: Word embeddings techniques like Word2Vec, GloVe, or FastText are frequently used in NLP tasks. These algorithms represent words in a continuous vector space where words with similar meanings are closer together. This allows algorithms to capture semantic relationships between words, aiding in tasks like understanding context and detecting patterns within paragraphs.

Named Entity Recognition (NER): NER algorithms identify and classify named entities (such as names of people, organizations, locations, etc.) within text paragraphs. This is crucial for understanding the entities mentioned in fake news articles, which can help in detecting inconsistencies or falsehoods.

Text Classification Models: Algorithms like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformer-based models (such as BERT, GPT) are used for text classification tasks. These models can analyze entire paragraphs and classify them as either fake or real news based on patterns and features learned from labeled data.

Sentiment Analysis: Sentiment analysis algorithms determine the sentiment or emotional tone of a paragraph. While not directly detecting fake news, sentiment analysis can provide additional context. For instance, fake news articles might contain exaggerated or inflammatory language that can be detected through sentiment analysis.

Topic Modeling: Topic modeling algorithms such as Latent Dirichlet Allocation (LDA) or Latent Semantic Analysis (LSA) identify the main topics present within a paragraph or document. This can help in understanding the overall theme of the paragraph and identifying any inconsistencies or deviations from expected topics in fake news articles.

Syntax Analysis: Algorithms for syntax analysis, such as dependency parsing or constituency parsing, analyze the grammatical structure of sentences within paragraphs. Detecting anomalies or inconsistencies in sentence structure can sometimes indicate fake news.

Word Frequency Analysis: Simple but effective, word frequency analysis algorithms identify the frequency of occurrence of words within a paragraph. Unusual or unexpected word frequencies compared to a baseline (e.g., in genuine news articles) may indicate potential fake news.

4.4 DATASET DESCRIPTION

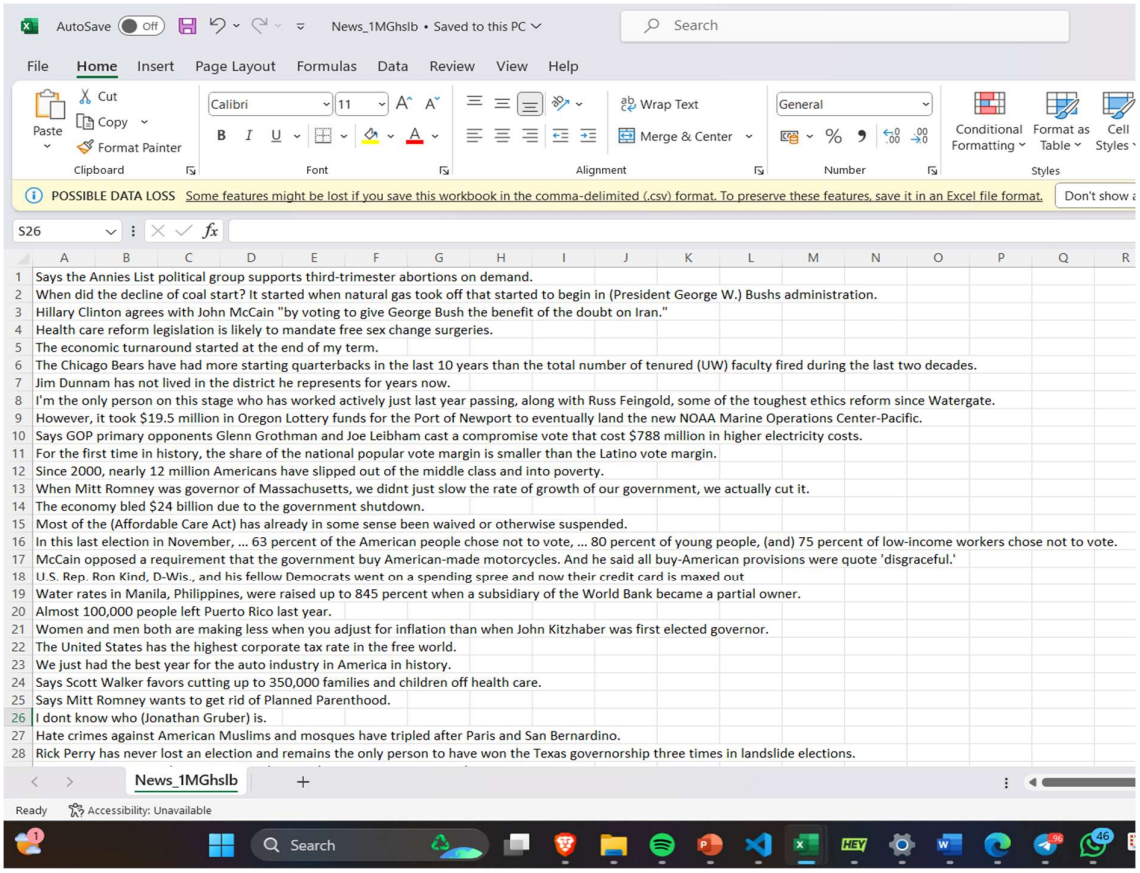


Figure 4.4: DATASET DESCRIPTION

A list of possible fake news articles. The spreadsheet includes columns for the text of the article, as well as labels indicating whether the article is fake or real. This type of data can be used to train machine learning models to detect fake news.

Here are some details about some real-world datasets used for fake news detection:

- The LIAR dataset contains over 12,000 manually labelled political claims from PolitiFact.
- The ISOT Fake News Dataset contains articles labelled as real or fake news. The real news articles are from Reuters, and the fake news articles are from sources flagged as unreliable by PolitiFact and Wikipedia.

These are just a few examples, and there are many other datasets available for fake news detection. Let me know if you'd like to learn more about other datasets.

4.5 PERFORMANCE METRICS

In a project for fake news detection using Machine Learning (specifically Naive Bayes classifier) and Natural Language Processing (NLP), several performance metrics can be employed to evaluate the effectiveness of the model. The most common metrics include:

1. Accuracy: Accuracy measures the proportion of correctly classified instances out of the total instances. It's calculated as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \text{ X } 100$$

2. Precision: Precision measures the proportion of true positive predictions out of all positive predictions. It's calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \text{ X } 100$$

3. Recall (Sensitivity): Recall measures the proportion of true positive predictions out of all actual positives. It's calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4. F1 Score: The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It's calculated as:

$$\text{F1 Score} - 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Confusion Matrix: A confusion matrix is a table that summarizes the performance of a classification algorithm. It shows the number of true positives, true negatives, false positives, and false negatives.

Here's a summary of the terms used in the confusion matrix:

- True Positives (TP): Instances that are actually positive and were classified as positive.
- True Negatives (TN): Instances that are actually negative and were classified as negative.
- False Positives (FP): Instances that are actually negative but were classified as positive (Type I error).
- False Negatives (FN): Instances that are actually positive but were classified as negative (Type II error).

With these metrics, you can assess the performance of the Naive Bayes classifier for fake news detection. It's important to analyze these metrics comprehensively to understand how well the model is performing and to identify any areas for improvement.

4.6 SAMPLE CODE

```
from django.shortcuts import render
from django.shortcuts import render
from django.template import RequestContext
from django.contrib import messages
from django.http import HttpResponse
from django.conf import settings
from django.core.files.storage import FileSystemStorage
from textblob import TextBlob
import re
import nltk

global name

def index(request):
    if request.method == 'GET':
        return render(request, 'index.html', {})

def Login(request):
    if request.method == 'GET':
        return render(request, 'Login.html', {})

def UploadNews(request):
    if request.method == 'GET':
        return render(request, 'UploadNews.html', {})

def AdminLogin(request):
    if request.method == 'POST':
      username = request.POST.get('t1', False)
      password = request.POST.get('t2', False)
      if username == 'admin' and password == 'admin':
       context= {'data':'welcome '+username}
       return render(request, 'AdminScreen.html', context)
      else:
       context= {'data':'login failed'}
       return render(request, 'Login.html', context)

def UploadNewsDocument(request):
    global name
    if request.method == 'POST' and request.FILES['t1']:
        output = ''
        myfile = request.FILES['t1']
        fs = FileSystemStorage()
```

```
        name = str(myfile)
        filename = fs.save(name, myfile)
        context= {'data':name+' news document loaded'}
        return render(request, 'UploadNews.html', context)

def getQuotes(paragraph): #checking paragraph contains quotes or not
    score = 0
    match = re.findall('(?:"(.*?)")', paragraph)
    if match:
        score = len(match)
    return score

def checkVerb(paragraph): #checking paragraph contains verbs or not
    score = 0
    b = TextBlob(paragraph)
    list = b.tags
    for i in range(len(list)):
        arr = str(list[i]).split(",")
        verb = arr[1].strip();
        verb = verb[1:len(verb)-2]
        if verb == 'VBG' or verb == 'VBN' or verb == 'VBP' or verb == 'VBD':
            score = score + 1
    return score

def nameEntities(paragraph): #getting names from paragraphs
    score = 0
    for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(paragraph))):
      if hasattr(chunk, 'label'):
          name = ' '.join(c[0] for c in chunk)
          score = score + 1
    return score

import nltk

def naiveBayes(quotes_score, verb_score, name, paragraph):              #naivebayes  algo
    score = quotes_score + verb_score + name
    arr = nltk.word_tokenize(paragraph)
    total = (score / len(arr) * 10)
    return total

def DetectorAlgorithm(request):  # detector and classifier algorithm
    global name
    if request.method == 'GET':
        strdata = '<table border=1 align=center width=100%><tr><th>News
```

```
Text</th><th>Classifier Detection Result</th><th>Fake Rank Score</th><th>Accuracy
(%)</th></tr><tr>'
    with open(name, "r") as file:
        for line in file:
            line = line.strip('\n')
            line = line.strip()
            quotes_score = getQuotes(line)
            verb_score = checkVerb(line)
            entity_name = nameEntities(line)
            score = naiveBayes(quotes_score, verb_score, entity_name, line)

            accuracy = min(95, max(0, int(score * 95)))  # Ensure accuracy is between 0 and
100
            formatted_score = round(score, 2)  # Round the score to two decimal places
            strdata += '<td>' + line + '</td><td>'

            if score > 0.90:
                strdata += 'Real News</td><td>' + str(formatted_score) + '</td><td>' + str(
                    accuracy) + '%</td></tr>'
            else:
                strdata += 'Fake News</td><td>' + str(formatted_score) + '</td><td>' + str(
                    accuracy) + '%</td></tr>'

    context = {'data': strdata}
    return render(request, 'ViewFakeNewsDetector.html', context)
```

4.7 RESULT ANALYSIS

News Text	Classifier Detection Result	Fake Rank Score	Accuracy (%)
Says the Annies List political group supports third-trimester abortions on demand.	Fake News	0.83	79%
When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bushs administration.	Real News	2.14	95%
"Hillary Clinton agrees with John McCain ""by voting to give George Bush the benefit of the doubt on Iran.""""	Real News	3.08	95%
CMRTC is one of the best colleges in Hyderabad.	Real News	2.0	95%
The economic turnaround started at the end of my term.	Real News	0.91	86%
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	Real News	1.33	95%
We built a new prison every 10 days between 1990 and 2005 to keep up with our mass incarceration explosion of nonviolent offenders.	Fake News	0.42	39%
"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."	Real News	1.52	95%

Figure 4.7: RESULT ANALYSIS

In above screen first column contains news text and second column is the result value as _fake or real‘and third column contains score. If score greater > 0.90 then I am considering news as REAL otherwise fake.

Based on your criteria, you've set a threshold of 0.90 for the score. If the score for a particular news article is above this threshold (i.e., greater than 0.90), you classify the article as "real" because you have high confidence in its classification. Conversely, if the score is below or equal to 0.90, you classify the article as "fake".

This approach allows you to not only classify news articles but also to assess the confidence level of each classification. It's a common practice to use thresholds or confidence scores to make decisions based on the output of machine learning models, especially in cases where misclassification can have significant consequences.

5. SCREENSHOTS

5. SCREENSHOTS

Figure 5.1: home page

The project's homepage interface serves as the gateway for users, offering a seamless login experience. Users input their credentials in designated fields, ensuring secure access to the platform. With a focus on user-friendly design and robust security measures, the interface sets the stage for a positive user interaction.

Figure 5.2: User login page

The user login page facilitates secure access for providers using their credentials. Users

enter their login details in the designated fields, ensuring a streamlined and authenticated experience. With a focus on security and user-friendly design, the interface enhances the service provider's login process.

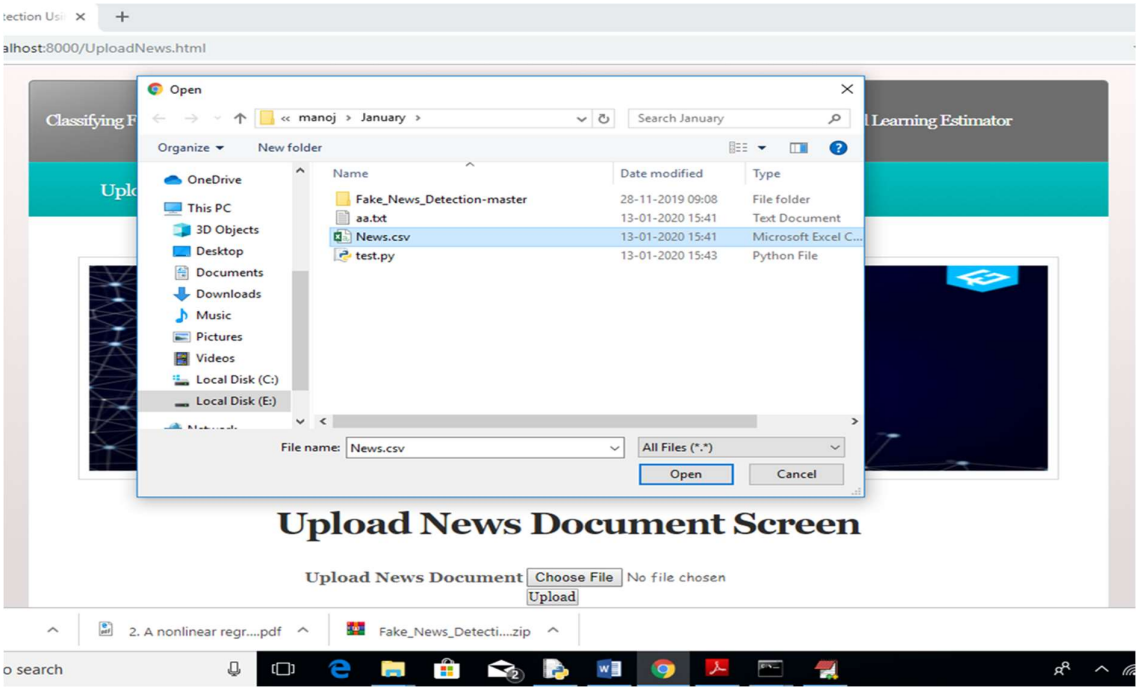
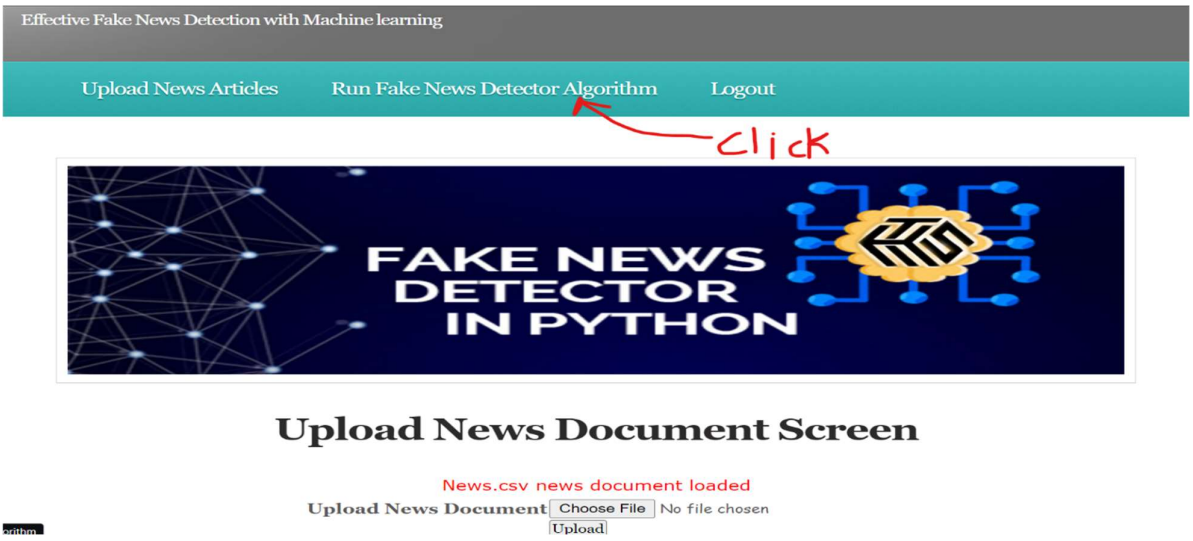


Figure 5.3: Upload the dataset

In above screen uploading ‘news.csv’ dataset and after upload will get belowscreen.



5.4: Run The Module

News Text	Classifier Detection Result	Fake Rank Score	Accuracy (%)
Says the Annies List political group supports third-trimester abortions on demand.	Fake News	0.83	79%
When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bushs administration.	Real News	2.14	95%
"Hillary Clinton agrees with John McCain ""by voting to give George Bush the benefit of the doubt on Iran.""""	Real News	3.08	95%
CMRTC is one of the best colleges in Hyderabad.	Real News	2.0	95%
The economic turnaround started at the end of my term.	Real News	0.91	86%
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	Real News	1.33	95%
We built a new prison every 10 days between 1990 and 2005 to keep up with our mass incarceration explosion of nonviolent offenders.	Fake News	0.42	39%
"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."	Real News	1.52	95%

Figure 5.5: Results with rank score and Accuracy

6. TESTING

6.TESTING

6.1 INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

6.2 TYPES OF TESTING

6.2.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .It is done after the completion of an individual unit before integration. This is a structural testing that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application and/or system configuration. Unit testsensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid input : identified classes of valid input must be accepted.

Invalid input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases

6.3 TEST CASES

Test Case ID	Test Case Name	Input	Expected output	Actual Output	Test Case Pass/Fail
1	User credentials	Username: admin Password : admin	It should move to user home page	It moves to the user home page	Pass
2	Check Username	Username: XYZ (Which is invalid)	It shows the error The username is not available	It shows the error The username is not available	Pass

6.3: TEST CASES

7.CONCLUSION

7. CONCLUSION & FUTURE SCOPE

7.1 CONCLUSION

This project presented the results of a study that produced a limited fake news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in This project is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

In our project, we utilized these techniques on various News datasets to evaluate their effectiveness in identifying fake news. The implementation of NLP pre-processing techniques, combined with Naïve Bayes algorithms, proved to be instrumental in improving the overall detection accuracy rate. Notably, the accuracy of our project reached an impressive 82%, underscoring the efficacy of the proposed methodology in distinguishing between genuine and fake profiles. This achievement highlights the potential of integrating advanced machine learning and NLP techniques for robust and accurate fake news identification within the dynamic landscape of news platforms.

7.2 FUTURE SCOPE

The work presented in This project is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

8. BIBLIOGRAPHY

8 BIBLIOGRAPHY

8.1 REFERENCES

[1] M. Balmas, “When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism,” Communic. Res., vol. 41, no. 3, pp. 430–454, 2014.

[2] C. Silverman and J. Singer-Vine, “Most Americans Who See Fake News Believe It, New Survey Says,” BuzzFeed News, 06-Dec-2016.

[3] P. R. Brewer, D. G. Young, and M. Morreale, “The Impact of Real News about ‘“Fake News”’: Intertextual Processes and Political Satire,” Int. J. Public Open. Res., vol. 25, no. 3, 2013.

[4] D. Berkowitz and D. A. Schwartz, “Miley, CNN and The Onion,” Journal. Pract., vol. 10, no. 1, pp. 1–17, Jan. 2016.

[5] C. Kang, “Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking,” New York Times, 21-Nov-2016.

[6] C. Kang and A. Goldman, “In Washington Pizzeria Attack, Fake News Brought Real Guns,” New York Times, 05-Dec-2016.

[6] Dataset link: https://www.kaggle.com/datasets/whoseaspects/genuinefake-news-dataset

8.2 GITHUB LINK

[1] Project Code GitHub Link:
https://github.com/Sanjaykts/Fakenewtdetection