FAKE NEWS DETECTION SYSTEM

MINIPROJECT REPORT

Submitted by

VASA CHANDANA [RA2011047010080]
RAJ MUKHERJEE [RA2011047010081]
BALLA GEETESH NIHAL [RA2011047010087]
SHRUTI IYENGAR [RA2011047010105]
BHAVYA YADAV [RA2011047010149]

Under the guidance of

Dr. MAHESHWARI A
(Guide Affiliation)
Assistant Professor

Department of Computational Intelligence

FACULTY OF ENGINEERING AND TECHNOLOGY

SCHOOL OF COMPUTING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Kancheepuram

MAY 2023

SRM INSTITUTEOF SCIENCEANDTECHNOLOGY

(UnderSection3ofUGCAct,1956)

BONAFIDECERTIFICATE

Certified that 18AIC305T_Inferential Statistics and Predictive Analytics titled "**FAKE NEWS DETECTION SYSTEM**" is the bonafide work of "**VASA CHANDANA [RA2011047010080], RAJ MUKHERJEE [RA2011047010081], BALLA GEETESH NIHAL [RA2011047010087], SHRUTI IYENGAR [RA2011047010105] and BHAVYA YADAV [RA2011047010149]"** who carried out the minor project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE
Faculty In-Charge
Dr. Maheshwari A
Assistant Professor
Department of Computational Intelligence,
SRM Institute of Science and Technology
Kattankulathur Campus, Chennai

HEAD OF THE DEPARTMENT
Dr. R Annie Uthra
Professor and Head ,
Department of Computational Intelligence,
SRM Institute of Science and Technology
Kattankulathur Campus, Chennai

ABSTRACT

Counterfeit news is the purposeful spread of deception or promulgation by conventional news media and through online media. Such news stories can come in numerous structures, including: accidental mistakes submitted by news aggregators, inside and out bogus stories, or the tales which are created to deceive and impact per user's assessment. Likewise, since bogus data can spread so quick, not just it can hurt individuals yet additionally can be hindering to tremendous enterprises and financial exchanges.

In the previous decade, we have seen dramatic development of data accessible on the web. It is getting difficult to interpret valid from the bogus. It is essential for us to know that how much of what we read on supposedly credible news site is trustworthy. False information can cause panic among citizens. Likewise lies can be utilized to control different people choices for casting a ballot bid or basically whatever else that can have enduring consequences. Bogus data spreads extremely quick, this is shown by the way that when one phony news site is brought down another immediately has its spot. Besides, it is fit for demolishing the equilibrium of the news environment. Political plans and control are one of the numerous intentions since counterfeit news is created.

Today, fortunately we have advances machine learning and Language Processing (NLP) instruments offer incredible guarantee for specialists to construct frameworks which could naturally identify counterfeit news. Hence it would be beneficial to discuss the methods of detecting online deceptions. Our undertaking can be essentially utilized by any media organization to foresee if the coursing news is phony. The interaction should be possible consequently without having people physically audit a huge number of information related articles.

Approaching the problem from a purely NLP perspective, it will be possible for us to predict whether the news is fake or real based only on its content. A significant piece of the objective is to think about and report the outcomes from numerous distinctive model executions, and present an investigation of the discoveries.

INDEX

LIST OF FIGURES

LIST OF TABLES

CHAPTER1

INTRODUCTION

In various venues, fake news is causing a variety of concerns, ranging from humorous articles to falsified news and planned government propaganda. A significant piece of the objective is to think about and report the outcomes from numerous distinctive model executions, and present an investigation of the discoveries.

The relevance of disinformation in American political discourse has received a lot of attention recently, especially in the aftermath of the presidential election in the United States. The phrase 'fake news' came to be used to designate factually erroneous and deceptive items that were published primarily for the goal of generating revenue through page views. This paper appears to have created a model that can reliably estimate the chances of a given article being false news.

After media attention, Facebook was at the Centre of much criticism. They have already included a tool that flags counterfeit news on a website when a client comes across such a page, and the company has expressed openly that they're chipping away at a programmed system to distinguish false stories, it is obviously a troublesome undertaking. Since counterfeit news shows up on the two finishes of the political range, a given calculation should be politically unprejudiced while as yet giving real news sources on the two closures of the range equivalent weight. Furthermore, it is a difficult question of legitimacy. Be that as it may, it is important to comprehend what counterfeit news is and what different methodologies are tackle this problem. It is important to concentrate how AI and the investigation of characteristic dialects permits us to identify bogus news.

There is a Kaggle contest called the "Bogus News Challenge," and Facebook is utilizing artificial intelligence to sift counterfeit reports through of clients' channels. Battling counterfeit news is a conventional book classification project with a basic proposition. Whether it is feasible for to make a model that can recognize "genuine" and "counterfeit" news or not? Thus, a proposed exertion on accumulating a dataset of both phony and authentic news and utilizing a random forest classifier to foster a model to sort an article as bogus or genuine dependent on its words and expressions.

The fundamental objective is to identify the bogus news, which is an old-style text rating issue with a straightforward proposition. It is important to foster a model equipped for recognizing "genuine" and "counterfeit" news.

CHAPTER 2

LITERATURE SURVEY

Title 1: Automatic deception detection: Methods for finding fake news

Authors: Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November).

Abstract: The main purpose of their study was to construct linguistic cue techniques, word approach, rhetorical structure and discourse analysis, network analysis methods and SVM classification. SVM classification techniques were provided. These are text-based models, which offer very little or no improvement on existing approaches.

Result: Random Forest requires numerous features and the survey showed highest accuracy achievement of 95%.

Title 2: Weakly supervised learning for fake news detection on Twitter

Authors: Helm Stetter, S., & Paulheim, H. (2018, August).

Abstract: Fake news identification on Twitter has been poorly controlled, and every Tweet/Post has been categorised as binary classification issue. The grades are based exclusively on the post/tweet source. Authors utilise twitter API, DMOZ and utilised techniques such as naïve bays, svm, XG boost and neural nets. The author uses data sets manually. The statistics reveal 15% fraudulent tweets, 45% actual tweets, remaining unresolved postings.

Result: Results suggest that the rest of the post was undecided: 15 percent bogus tweets, 45 percent real tweets.

Title 3: Automatic Online Fake News Detection Combining Content and Social Signals.

Authors: Ballarin, G., DiPierro, M., & de Alfaro, L. (2018, May).

Abstract: Facebook messenger's chatbot implemented it. Three various datasets of Facebook Italian news posts were used. The Boolean crowd sourcing algorithms were implementable, both for content based approaches with social and content signals.

Result: Three various datasets of Facebook Italian news posts were used. The Boolish crowdsourcing algorithms are used for both contents- based approaches using social and content signals.

Some more Literature Surveys:

No.	Title	Author	Year	Findings	Limitations
1	"Fake News Detection on Social Media: A Data Mining Perspective"	Shu et al.	2017	Proposed a fake news detection framework that combines linguistic and network features.	Did not evaluate the framework on a large scale or in real-world scenarios.
2	"A Survey of Fake News Detection Tools: Techniques, Methods, and Opportunities"	Vosoughi et al.	2018	Reviewed various techniques and methods used for fake news detection, including content-based, network-based, and hybrid approaches.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
3	"Fake News Detection using Machine Learning: A Systematic Literature Review"	Gilda et al.	2019	Conducted a systematic review of existing fake news detection systems and found that machine learning-based approaches are the most effective.	Did not provide a comparative evaluation of different machine learning-based approaches.
4	"A Deep Learning Approach to Identifying Misinformation on Social Media"	Wang et al.	2019	Proposed a deep learning-based approach that combines both linguistic and visual features for detecting fake news.	Did not evaluate the approach on a large scale or in real-world scenarios.
5	"Fake News Detection on Social Media: An Information Retrieval Perspective"	Kim et al.	2020	Proposed an information retrieval-based approach that considers the relevance of news articles and the trustworthiness of their sources for fake news detection.	Did not evaluate the approach on a large scale or in real-world scenarios.
6	"A Survey of Fake News Detection: Fundamental Concepts, Methodologies, and Opportunities"	Wu et al.	2020	Reviewed various fundamental concepts and methodologies used for fake news detection, including linguistic, network, and visual-based approaches.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
7	"Fake News Detection Using Machine Learning Techniques: A Review"	Uysal et al.	2020	Conducted a comprehensive review of existing fake news detection systems that use machine learning techniques and identified various challenges and opportunities for future research.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.

8	"A Review of Fake News Detection Methods: From Classical Machine Learning to Deep Learning and Natural Language Processing"	Abbasi et al.	2020	Reviewed various fake news detection methods, including classical machine learning, deep learning, and natural language processing-based approaches, and identified their strengths and weaknesses.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
9	"A Survey of Fake News Detection Methods: From Handcrafted Features to Deep Learning Models"	Su et al.	2020	Reviewed various fake news detection methods, from handcrafted features to deep learning models, and identified the advantages and limitations of each approach.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
10	"Fake News Detection: A Systematic Literature Review"	Venkatesan et al.	2020	Conducted a systematic literature review of fake news detection and identified various challenges and opportunities for future research, including the need for more comprehensive datasets and robust evaluation metrics.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
11	"Fake News Detection: A Comprehensive Review of Existing Methods and Techniques"	Singh et al.	2021	Conducted a comprehensive review of existing fake news detection methods and techniques, including linguistic, network, and visual-based approaches, and identified the advantages and limitations of each approach.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
12	"A Systematic Review of Fake News Detection Techniques: State-of-the-Art and Future Research Directions"	Mohanta et al.	2021	Conducted a systematic review of fake news detection techniques and identified the challenges and opportunities for future research, including the need for more robust and scalable algorithms.	Did not propose any new detection techniques or evaluate the effectiveness of existing ones.
13	"Fake News Detection: A Systematic Literature Review and a Proposal for a Novel Approach"	Spina et al.	2021	Conducted a systematic literature review of fake news detection and proposed a novel approach based on sentiment analysis and machine learning techniques.	Did not evaluate the effectiveness of the proposed approach on a large scale or in real-world scenarios.
14	"Fake News Detection Using Machine Learning Techniques: A	Singh et al.	2021	Conducted a comprehensive review of existing fake news detection systems that use machine learning techniques and identified the challenges	Did not propose any new detection techniques or evaluate the

	Comprehensive Review"			and opportunities for future research.	effectiveness of existing ones.
15	"A Comprehensive Survey of Fake News Detection: Techniques and Challenges"	Parab et al.	2021	Reviewed various fake news detection techniques, including linguistic, network, and visual-based approaches, and identified the challenges and opportunities for future research.	Did not propose any new detection techniques or evaluate

CHAPTER 3

PROPOSED WORK

Identify the problem and define the scope of the project: The first step in developing a fake news detection system is to clearly define the problem and establish the scope of the project. This involves identifying the types of fake news to be detected, the sources of the fake news, and the intended audience.

Gather a dataset: The second step is to gather a dataset of fake and real news articles. The dataset should be diverse, well-balanced, and representative of the types of fake news to be detected.

Preprocess the data: The third step is to preprocess the data. This involves cleaning and formatting the data to ensure that it is consistent and in a format that can be easily analyzed.

Feature extraction: The fourth step is to extract relevant features from the data. This involves using natural language processing (NLP) techniques to identify features such as the presence of specific words, the tone of the article, and the writing style.

Model selection: The fifth step is to select an appropriate machine learning algorithm to use for the classification of news articles. This may involve testing different algorithms to determine which one performs best on the dataset.

Model training: The sixth step is to train the selected machine learning model on the dataset. This involves feeding the preprocessed data into the model and adjusting the model parameters to optimize its performance.

Model evaluation: The seventh step is to evaluate the performance of the trained model. This involves using a separate dataset of news articles to test the accuracy of the model's predictions.

Refinement and optimization: The eighth step is to refine and optimize the model to improve its accuracy. This may involve adjusting the feature extraction process, selecting a different machine learning algorithm, or tuning the model parameters.

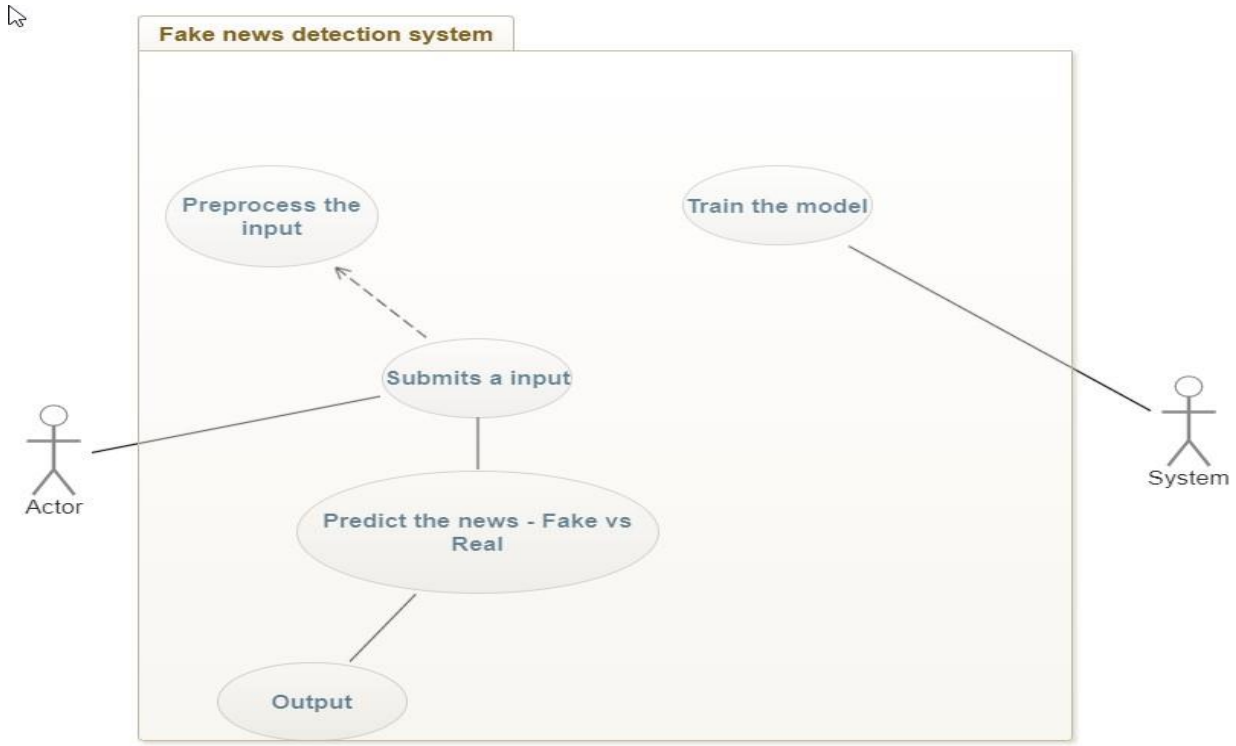
Deployment: The final step is to deploy the fake news detection system. This may involve integrating the model into a web-based or mobile application, or incorporating it into an existing news platform to flag potentially fake articles for review.

In summary, the development of a fake news detection system involves identifying the problem, gathering and preprocessing data, extracting features, selecting and training a machine learning model, evaluating its performance, refining and optimizing the model, and finally deploying the system.

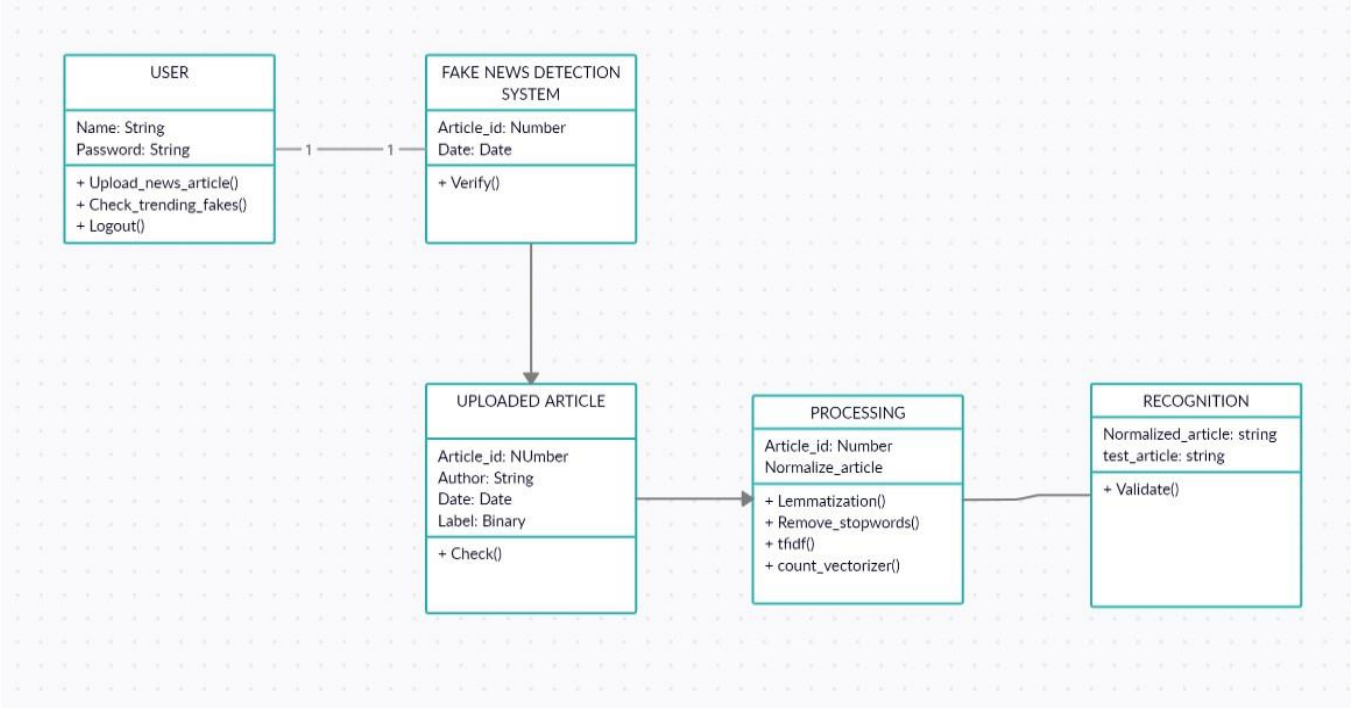
USE CASE DIAGRAM:

Scenario: User Interaction

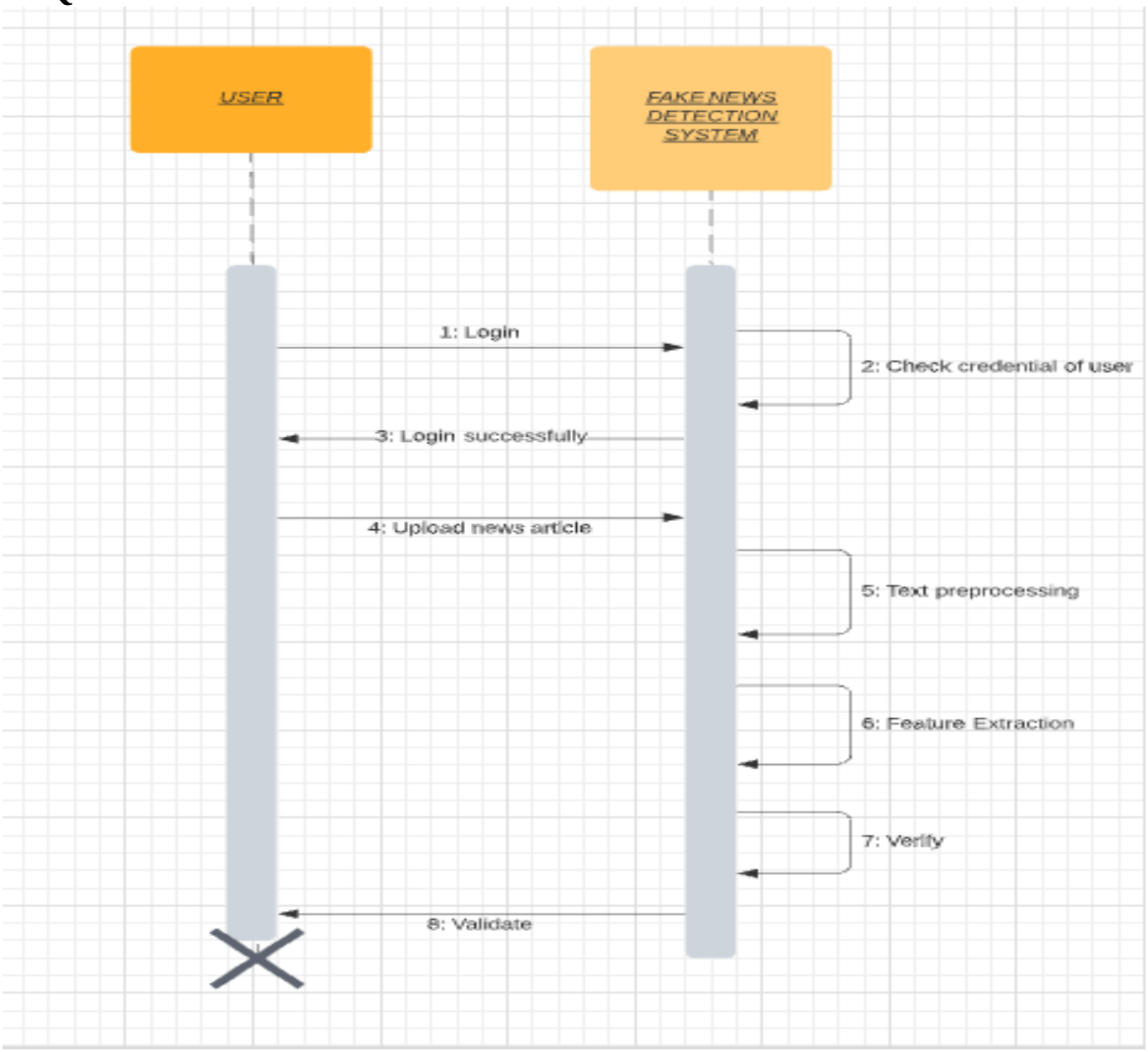
Actors: An actor is an individual, entity, or external structure that plays a role with our structure in one or more encounters (actors are usually drawn as UML Stick Figures Use case diagrams).



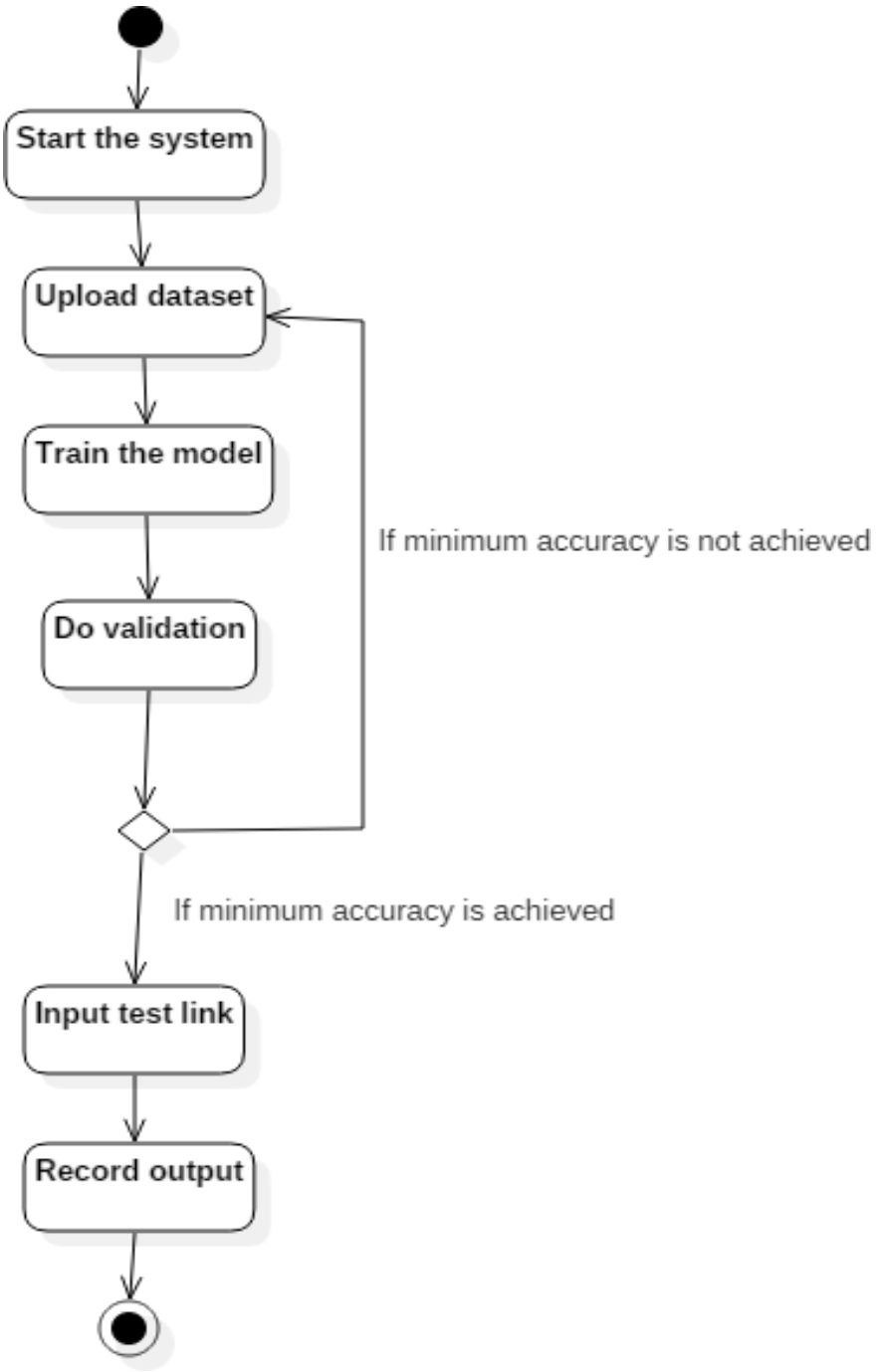
CLASS DIAGRAM:



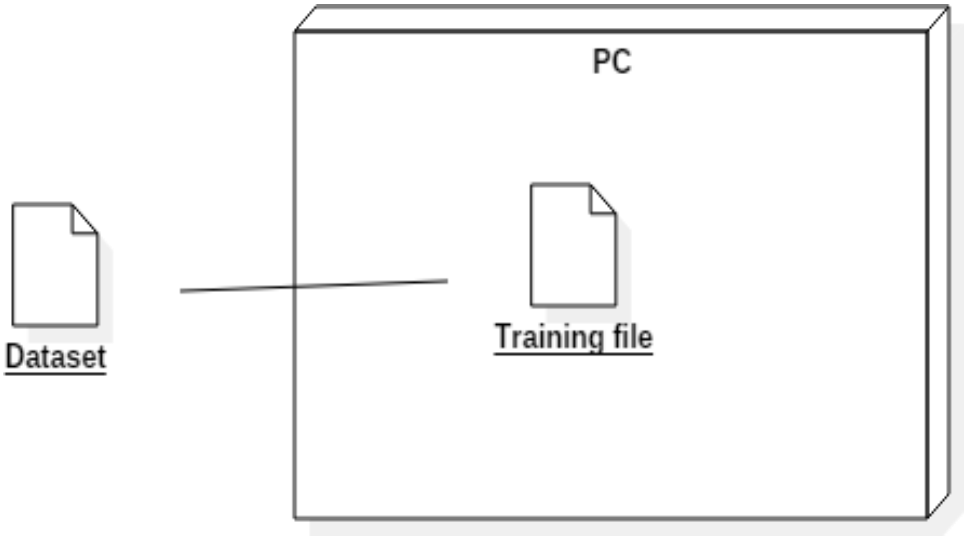
SEQUENCE DIAGRAM:



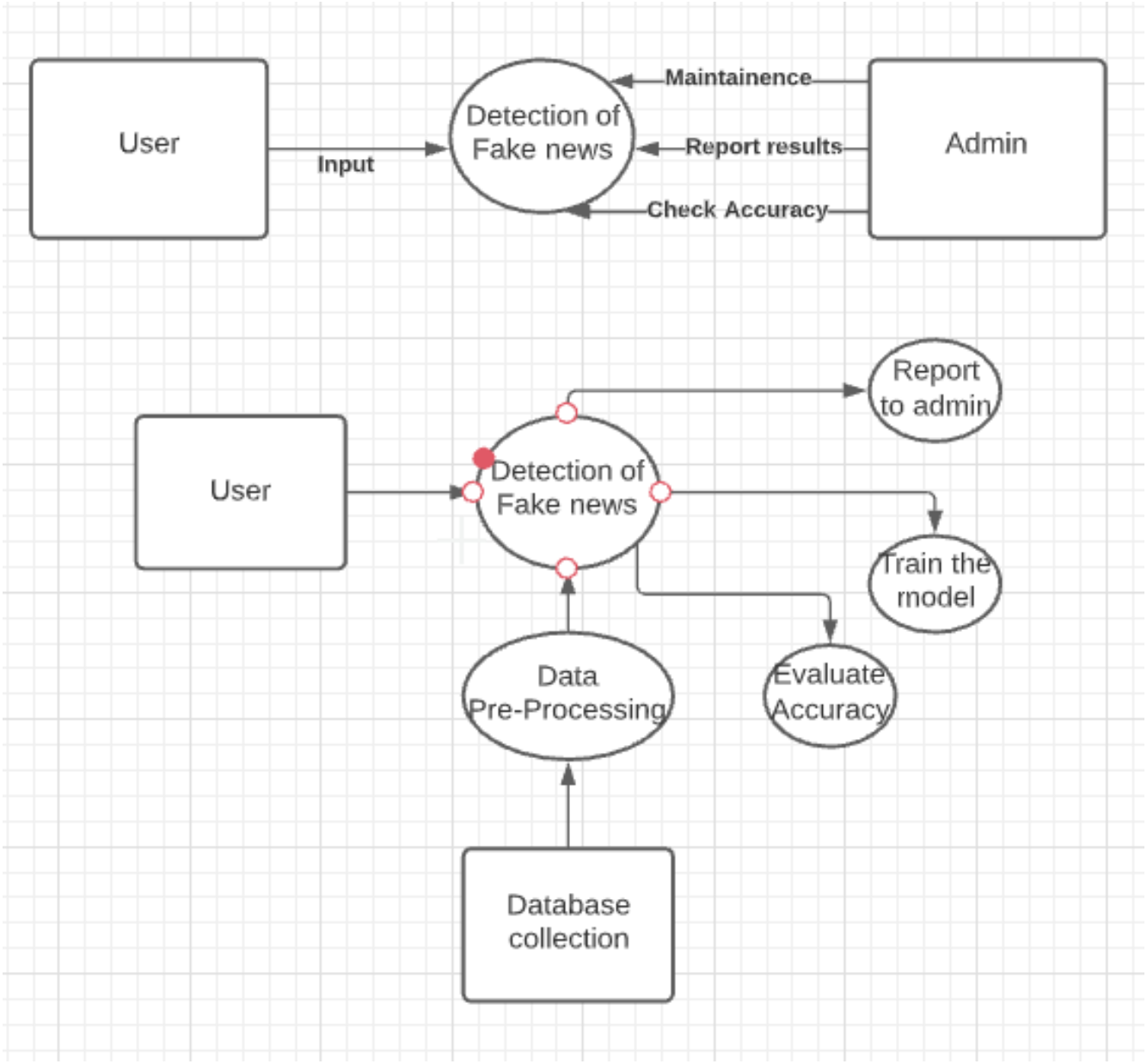
STATE/ACTIVITY DIAGRAM:



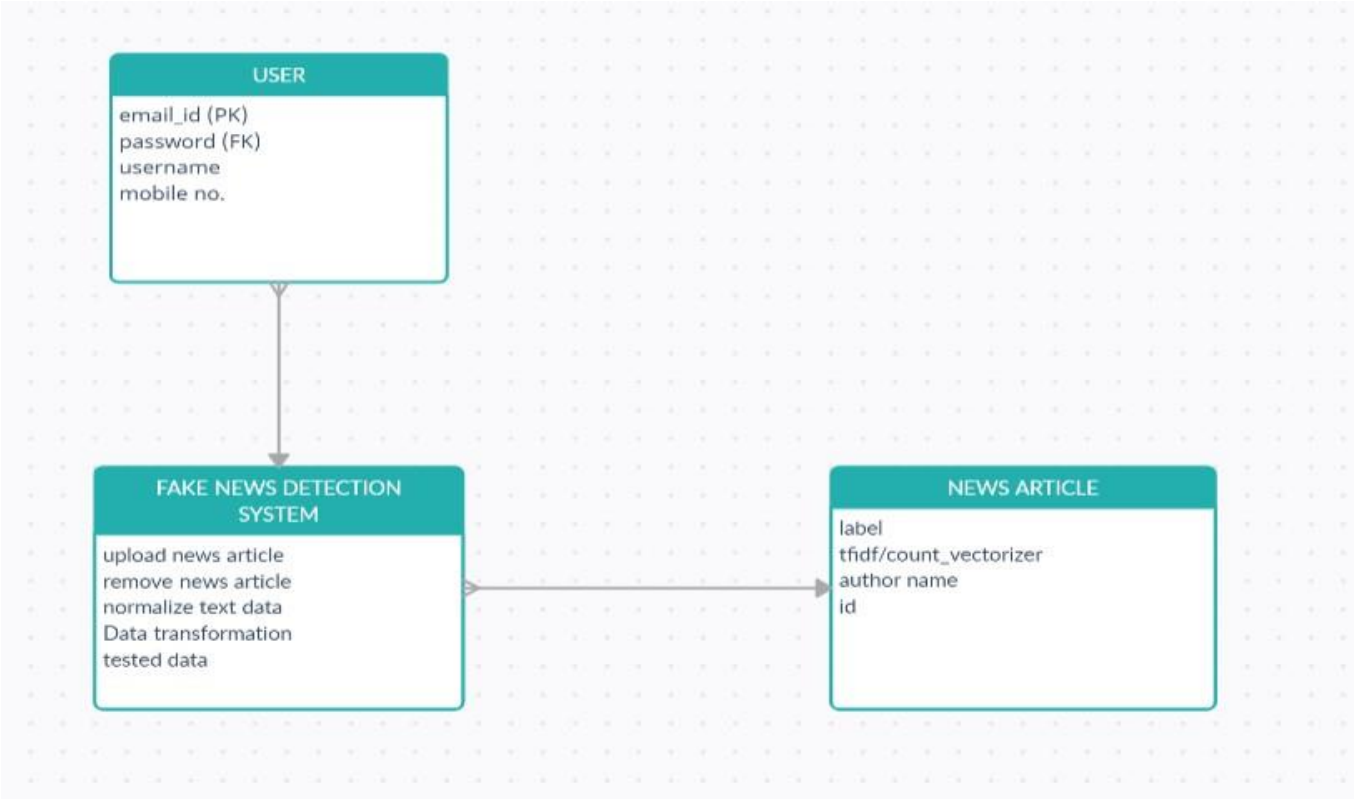
DEPLOYMENT DIAGRAM:



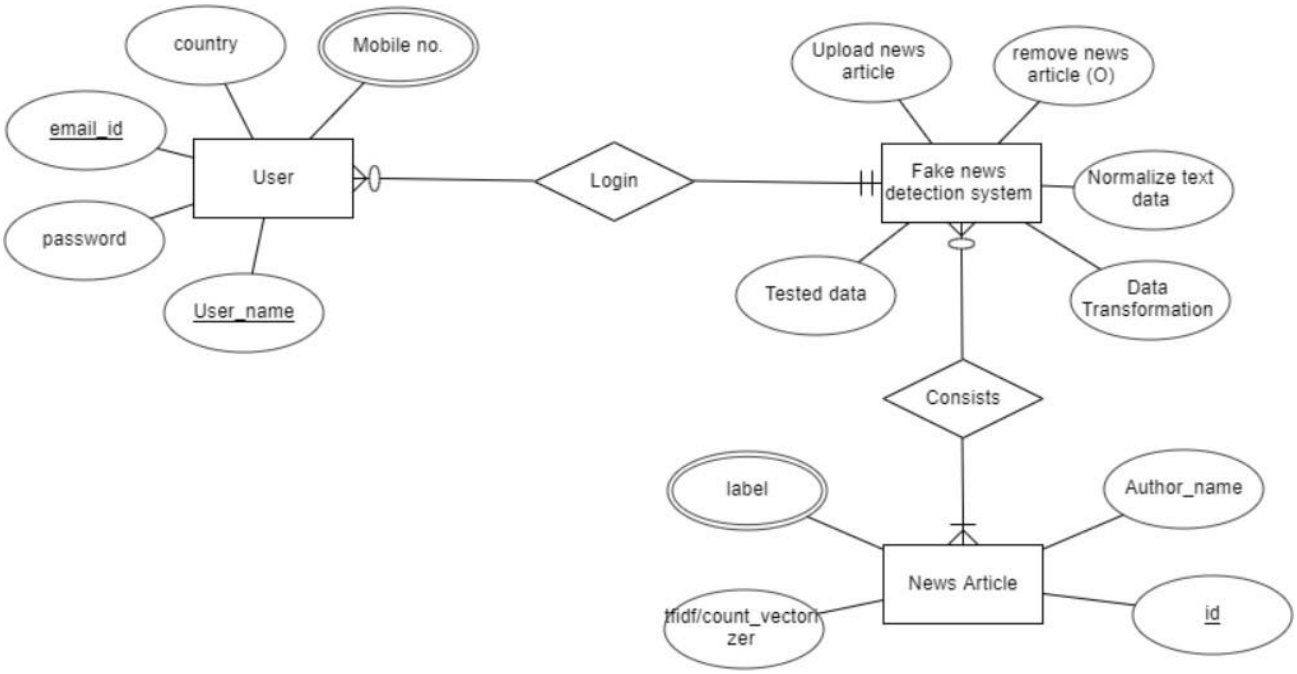
DATA FLOW DIAGRAM:



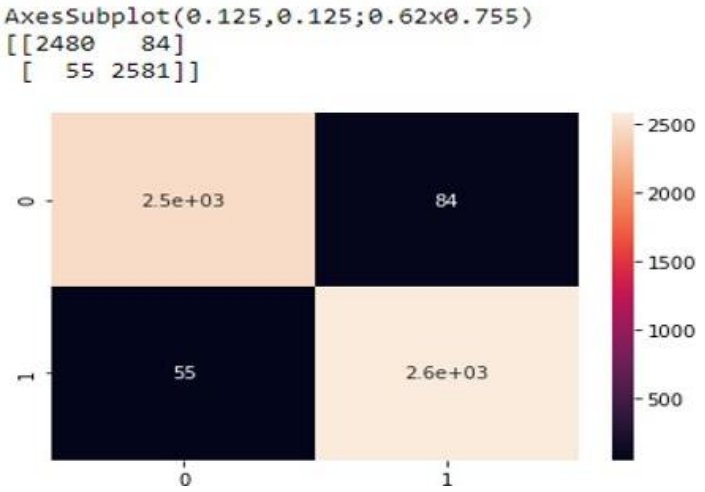
RELATIONAL DIAGRAM:



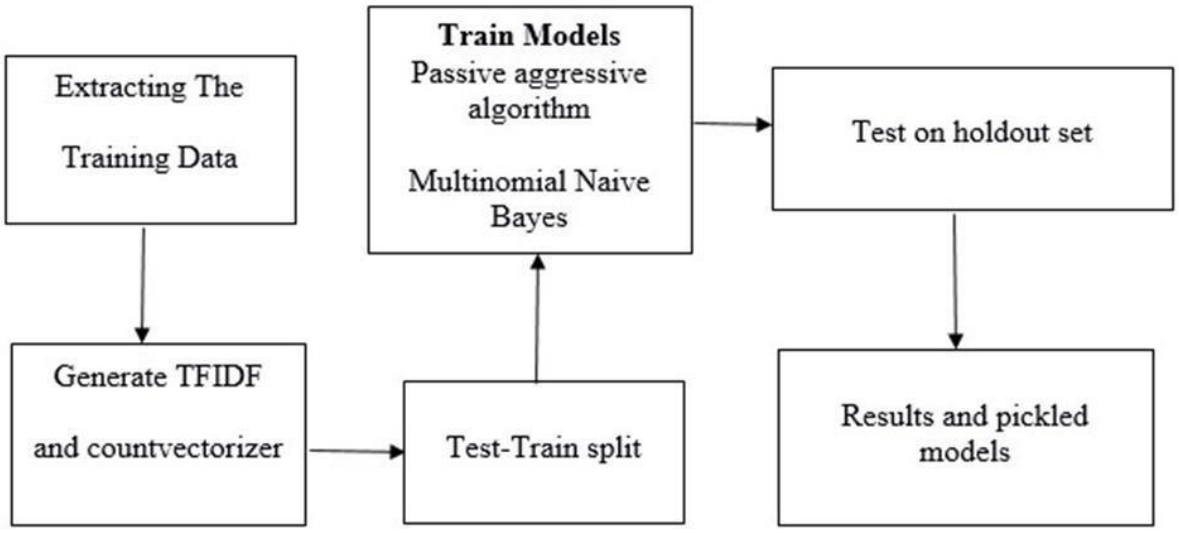
DATABASE DESIGN:



CONFUSION MATRIX: (With c-value=3.0)



SYSTEM ARCHITECTURE:



RISK ANALYSIS:

For the purpose of our project we have the following three major risks along with their counter measures exercised.

RISK	Counter-Measure
I. Underfitting	An AI model that is underfit is unsatisfactory, as confirmed by its horrible showing on the preparation information. It can't sum up to new information or model the preparation information. The arrangement is to go on and explore different avenues regarding diverse AI methods.
II. Overfitting	At the point when a model learns the data and clamor in the preparation information to the point that it debases the model's exhibition on new information, this is known as overfitting. Cross approval is an effective and gainful countermeasure against such a risk.
III. Sarcasm	Sarcasm detection is a fairly specialized study subject in NLP, a type of sentiment analysis where the focus is on recognizing sarcasm rather than recognizing a sentiment throughout the entire spectrum. When it comes to making accurate forecasts, it's still a difficulty.

Along with this there is always risk of other factors such as:

- User input is invalid
- Quality Issues
- Hardware Issues

VERIFICATION AND VALIDATION:

Unit Testing:

TEST ID	TEST ACTIONS	INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	PASS/FAIL
T1	Check the integrity of the dataset	Access the dataset and check the attributes of the slices	Access to the dataset and all its attributes	Success	Pass
T2	Checking for null values and removing them	Testing for null values	Null values replaced by ‘’	Success	Pass
T3	Creating a word cloud	Dataset of true events	Word cloud of essential words	Success	Pass
T4	Implementing count vectorizer	Train dataset	Vectorized table of term frequency	Success	Pass
T5	Implementing Tf idf vectorizer	Count vectorized data	Essential words vector	Success	Pass
T6	Passing vectorized data into Logistic regression classifier	Pre-processed vector	Successful prediction	Success	Pass
T7	Passing vectorized data into naïve bayes classifier	Pre-processed data vector	Successful prediction	Success	Pass

Integration Testing:

TEST DESCRIPTION	TEST STEPS	TEST DATA	EXPECTED RESULT	ACTUAL RESULT
We integrated the unit tested module and tested the behavior as a complete unit.	1) Integration of code (from units) 2) Running the code: • Input the data sets • Cleaning and lemmatization of the data • Tfidf vectorization • Passing through different classifiers to get optimal accuracy	2 datasets, namely, test and train data sets with 20,800 entries were incorporated into the model.	Efficiencies expected were as follows: Support Vector Classifier With C=1.0 Accuracy =95 % With C=2.0 Accuracy= 96% With C=10.0 Accuracy= 97%	Efficiencies obtained were as follows: Support Vector Classifier With C=1.0 Accuracy = 95.32 % With C=2.0 Accuracy= 97.4815% With C=10.0 Accuracy= 97.4816%

MC CALL'S QUALITY FACTOR:

QUALITY FACTOR	DEFINITION	JUSTIFICATION
Correctness	The extent to which a software meets its specifications and achieves the user's mission objectives. A software system is considered to be accurate if it meets all functional criteria.	In our project our sole objective was to classify fake news and real news in order to stop the spread the spread of misinformation. We can accomplish this with a precision of 97% utilizing SVM classifier
Reliability	The degree to which a programme can be trusted to fulfil its intended purpose with the needed precision is referred to as its reliability. Reliability is a consumer impression, and faulty software might nevertheless be regarded as dependable.	We are able to perform our intended function with a precision score of 98% and 97% respectively for values labelled true and false using SVM classifiers.
Efficiency	The amount to which a software system uses resources such as processing power, memory, disc space, network bandwidth, and energy is referred to as efficiency. A software should ideally use as few computational resources as feasible.	Our project is operational on the following minimal hardware requirements: Intel core i5 Ram: 8GB Hard disk space: 100GB Which is relatively small compared to other deployed machine/deep learning modules.
Integrity	The capacity of a system to withstand security threats is referred to as its integrity. In other terms, integrity relates to the degree to which unauthorized people or programs may access software or data.	Jupyter notebook provides provision for setting authorization levels. A unique password can be set which restricts access to the jupyter notebook server.
Portability	The amount of effort necessary to move a programme from one hardware and/or software environment to another.	Jupyter notebook and anaconda navigator are open source. Which means that anybody can share .ipynb files and run it on their respective workstations.

Reusability	To what extent portions of a software system in different applications may be utilised Reusability refers to the ability of a large piece of one product to be reused, maybe with minimal adjustments, in another product.	Our project deals with NLP so the data pre-processing steps like tokenisation and lemmatization remains the same when it comes to implementation. Those parts of the code and extracted and used for other NLP applications.

CHAPTER 4

MODULE AND ALGORITHM DESCRIPTION

A. Kaggle Data Collection

- In this module, we have collected the dataset from Kaggle which have around 25117 rows of data with 5 fields named id, title, author, text and label in train.csv and 5881 rows of data with 5 fields named id, title, author and text.

B. Importing Dependencies and Data Pre-processing

- We use the following Libraries:

  Numpy, Pandas, sklearn, Stopwords, PorterStemmer, train_test_split, LogisticRegression, accuracy_score
- Text Cleaning: The collected data needs to be cleaned to remove any noise or irrelevant information. This can include removing HTML tags, special characters, punctuations, and stop words.
- Text Normalization: Text normalization involves converting the text data to a standard format. This can include converting all text to lowercase, removing numbers, and expanding contractions.
- Tokenization: Tokenization involves breaking the text data into individual words or tokens. This is an important step for feature extraction and analysis.
- Stemming/Lemmatization: Stemming and lemmatization techniques can be used to reduce the words to their base form. This can help reduce the dimensionality of the data and improve feature extraction.
- Feature Extraction: Once the data has been preprocessed, you can extract features from the text data. Commonly used features for fake news detection can include linguistic features such as word frequency, sentiment analysis, or semantic similarity.
- Data Labeling: The preprocessed data needs to be labeled as either real or fake news to train the machine learning models.

C. Stemming

- In a fake news detection system project, the stemming process can be integrated into the data preprocessing module to improve the accuracy of feature extraction.
- In this project, we are using PorterStemming.

D. Logistic Regression

- In a fake news detection system project, logistic regression can be used in the classification stage after the data pre-processing and feature extraction stages. The system can be trained on a labelled dataset to learn the relationship between the input variables and the binary outcome.
- The experimental results can showcase the performance of the fake news detection system using logistic regression by comparing the results with other machine learning models.

E. Evaluation

- We split our dataset into Training and Testing sets to evaluate the systems performance on unseen data.
- In this project, we measure the accuracy score of the predicted data. These evaluation results can be helpful in providing valuable insights.

F. Making a Predictive System

- The predictive system can be designed to automatically classify news articles as fake or real based on a set of input features such as word frequency, sentiment, or readability scores. The system can be trained on a labeled dataset to learn the relationship between the input features and the binary outcome.
- Once the system is trained, it can be used to predict the outcome of new, unseen news articles.

G. Classifier Evaluation Metrics

- **Accuracy:** The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

CHAPTER 5

RESULT AND CONCLUSION

- In conclusion, this fake news detection system project using logistic regression demonstrated that machine learning techniques can be used to effectively detect fake news articles. The project involved several stages, including data collection, pre-processing, feature extraction, model training, and evaluation. The logistic regression algorithm was used to model the probability of a binary outcome based on a set of input features. The evaluation results showed that the system achieved high accuracy.
- However, the system had limitations in handling nonlinear relationships between the input variables and the outcome. Overall, the project demonstrated the potential of using machine learning techniques in the development of effective fake news detection systems. Further research can focus on improving the system's performance, incorporating more advanced machine learning algorithms, and considering the impact of external factors such as the source of the news article and the timeliness of the detection.

REFERENCES

- **IEEE Conference 2019:** - Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation.

- **IEEE Conference, 2019** 1st International Conference on Advances in Information Technology: - Fake News Detection Using Deep Learning Techniques.

- **IEEE Conference 2019**: - Fake News Detection Using Machine Learning approaches: A systematic Review.

- **2018 4th International Conference on Computing Communication and Automation (ICCCA)**: - Fake News Detection Using A Deep Neural Network.

APPENDIX

1.1 OUTPUT SCREENSHOT

Screenshots of Model Training and Output:

Importing the Dependencies

```
[2]  import numpy as np
     import pandas as pd
     import re
     from nltk.corpus import stopwords
     from nltk.stem.porter import PorterStemmer
     from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
```

```
[3]  import nltk
     nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

```
[4]  # printing the stopwords in English
     print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
```

Data Pre-processing

```
[5] # loading the dataset to a pandas DataFrame
    news_dataset = pd.read_csv('/content/train.csv')
```

```
[6] news_dataset.shape
```

```
(20800, 5)
```

```
[7] # print the first 5 rows of the dataframe
    news_dataset.head()
```

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

```
[8] # counting the number of missing values in the dataset
    news_dataset.isnull().sum()
```

```
id          0
title     558
author   1957
text       39
label       0
dtype: int64
```

```
[9] # replacing the null values with empty string
    news_dataset = news_dataset.fillna('')
```

```
[10] # merging the author name and news title
     news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']
```

```
[11] print(news_dataset['content'])
```

```
0        Darrell Lucus House Dem Aide: We Didn't Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                               ...
20795    Jerome Hudson Rapper T.I.: Trump a 'Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799                David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

```
[12] # separating the data & label
     X = news_dataset.drop(columns='label', axis=1)
     Y = news_dataset['label']
```

```
print(X)
print(Y)
```

```
          id                                              title  \
0          0  House Dem Aide: We Didn’t Even See Comey’s Let...
1          1  FLYNN: Hillary Clinton, Big Woman on Campus - ...
2          2                  Why the Truth Might Get You Fired
3          3  15 Civilians Killed In Single US Airstrike Hav...
4          4  Iranian woman jailed for fictional unpublished...
...      ...                                                ...
20795  20795  Rapper T.I.: Trump a ’Poster Child For White S...
20796  20796  N.F.L. Playoffs: Schedule, Matchups and Odds -...
20797  20797  Macy’s Is Said to Receive Takeover Approach by...
20798  20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799  20799                          What Keeps the F-35 Alive

                                      author  \
0                                Darrell Lucus
1                              Daniel J. Flynn
2                           Consortiumnews.com
3                              Jessica Purkiss
4                               Howard Portnoy
...                                        ...
20795                           Jerome Hudson
20796                         Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                             Alex Ansary
20799                           David Swanson

                                                    text  \
0      House Dem Aide: We Didn’t Even See Comey’s Let...
1      Ever get the feeling your life circles the rou...
2      Why the Truth Might Get You Fired October 29, ...
3      Videos 15 Civilians Killed In Single US Airstr...
4      Print \nAn Iranian woman has been sentenced to...
...                                                  ...
20795  Rapper T. I. unloaded on black celebrities who...
20796  When the Green Bay Packers lost to the Washing...
20797  The Macy’s of today grew from the union of sev...
20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799    David Swanson is an author, activist, journa...
```

```
                                                    content
0      Darrell Lucus House Dem Aide: We Didn’t Even S...
1      Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2      Consortiumnews.com Why the Truth Might Get You...
3      Jessica Purkiss 15 Civilians Killed In Single ...
4      Howard Portnoy Iranian woman jailed for fictio...
...                                                  ...
20795  Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796  Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797  Michael J. de la Merced and Rachel Abrams Macy...
20798  Alex Ansary NATO, Russia To Hold Parallel Exer...
20799            David Swanson What Keeps the F-35 Alive

[20800 rows x 5 columns]
0        1
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
20799    1
Name: label, Length: 20800, dtype: int64
```

Stemming:

Stemming is the process of reducing a word to its Root word

example: actor, actress, acting --> act

```
[14] port_stem = PorterStemmer()
```

```
[15] def stemming(content):
         stemmed_content = re.sub('[^a-zA-Z]',' ',content)
         stemmed_content = stemmed_content.lower()
         stemmed_content = stemmed_content.split()
         stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
         stemmed_content = ' '.join(stemmed_content)
         return stemmed_content
```

```
[16] news_dataset['content'] = news_dataset['content'].apply(stemming)
```

```
[17] print(news_dataset['content'])
```

```
0        darrel lucu hous dem aid even see comey letter...
1        daniel j flynn flynn hillari clinton big woman...
2                   consortiumnew com truth might get fire
3        jessica purkiss civilian kill singl us airstri...
4        howard portnoy iranian woman jail fiction unpu...
                               ...
20795    jerom hudson rapper trump poster child white s...
20796    benjamin hoffman n f l playoff schedul matchup...
20797    michael j de la merc rachel abram maci said re...
20798    alex ansari nato russia hold parallel exercis ...
20799                            david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

```
[18] #separating the data and label
     X = news_dataset['content'].values
     Y = news_dataset['label'].values
```

```
[19] print(X)
```

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' ...
 'michael j de la merc rachel abram maci said receiv takeov approach hudson bay new york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

```
[20] print(Y)
```

```
[1 0 1 ... 0 1 1]
```

```
[21] Y.shape
```

```
(20800,)
```

```
[22] # converting the textual data to numerical data
     vectorizer = TfidfVectorizer()
     vectorizer.fit(X)

     X = vectorizer.transform(X)
```

```
[23] print(X)
```

```
  (0, 15686)    0.28485063562728646
  (0, 13473)    0.2565896679337957
  (0, 8909)     0.3635963806326075
  (0, 8630)     0.29212514087043684
  (0, 7692)     0.24785219520671603
  (0, 7005)     0.21874169089359144
  (0, 4973)     0.233316966909351
  (0, 3792)     0.2705332480845492
  (0, 3600)     0.3598939188262559
  (0, 2959)     0.2468450128533713
  (0, 2483)     0.3676519686797209
  (0, 267)      0.27010124977708766
  (1, 16799)    0.30071745655510157
  (1, 6816)     0.1904660198296849
  (1, 5503)     0.7143299355715573
  (1, 3568)     0.26373768806048464
  (1, 2813)     0.19094574062359204
  (1, 2223)     0.3827320386859759
  (1, 1894)     0.15521974226349364
  (1, 1497)     0.2939891562094648
  (2, 15611)    0.41544962664721613
  (2, 9620)     0.49351492943649944
  (2, 5968)     0.3474613386728292
  (2, 5389)     0.3866530551182615
  (2, 3103)     0.46097489583229645
  :     :
  (20797, 13122)        0.2482526352197606
  (20797, 12344)        0.27263457663336677
  (20797, 12138)        0.24778257724396507
  (20797, 10306)        0.08038079000566466
  (20797, 9588) 0.174553480255222
  (20797, 9518) 0.2954204003420313
  (20797, 8988) 0.36160868928090795
  (20797, 8364) 0.22322585870464118
  (20797, 7042) 0.21799048897828688
  (20797, 3643) 0.21155500613623743
  (20797, 1287) 0.33538056804139865
  (20797, 699)  0.30685846079762347
  (20797, 43)   0.29710241860700626
  (20798, 13046)        0.22363267488270608
  (20798, 11052)        0.4460515589182236
  (20798, 10177)        0.3192496370187028
  (20798, 6889) 0.32496285694299426
  (20798, 5032) 0.4083701450239529
  (20798, 1125) 0.4460515589182236
  (20798, 588)  0.3112141524638974
  (20798, 350)  0.28446937819072576
  (20799, 14852)        0.5677577267055112
  (20799, 8036) 0.45983893273780013
  (20799, 3623) 0.37927626273066584
  (20799, 377)  0.5677577267055112
```

Splitting the dataset to training & test data

```
[24] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify=Y, random_state=2)
```

Training the Model: Logistic Regression

```
[25] model = LogisticRegression()
```

```
[26] model.fit(X_train, Y_train)
```

```
LogisticRegression()
```

Evaluation

accuracy score

```
[27] # accuracy score on the training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

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

```
Accuracy score of the training data :  0.9865985576923076
```

```
[29] # accuracy score on the test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

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

```
Accuracy score of the test data :  0.9790865384615385
```

Making a Predictive System

```
[31] X_new = X_test[3]

     prediction = model.predict(X_new)
     print(prediction)

     if (prediction[0]==0):
       print('The news is Real')
     else:
       print('The news is Fake')
```

```
[0]
The news is Real
```

```
[32] print(Y_test[3])
```

```
0
```