Fake News Detection

Group Members:
Sidharth Nair-16010120032
Rahi Patil-16010120038
Manan Sayar-16010120045

Project Mentor:Mrs Smita Sankhe

Chapter 1: Introduction

1.1 Background/Motivation

Now-a-days no one reads newspapers. Everyone is dependent on the virtual world to get the news which is happening around. But one cannot trust the source of the information which is passed in this virtual world and believes whatever information they get. This increases the risk of fake news and misconceptions among the real world. To reduce the risk of such fake news and misconceptions we are proposing a Fake News Detection System using Machine Learning models which will automatically detect the wrong information being passed and would reduce the fake rumors about something and also reduce the misconception among the real world.

1.2 Problem Statement

Develop a methodology for identifying fake news by leveraging online newspapers and social media platforms through text scraping. The accuracy of the proposed approach will be evaluated using several machine learning algorithms. This research aims to contribute to the ongoing efforts to combat the dissemination of misleading information by providing an effective and reliable means of detecting fake news.

1.3 Scope

Our project's focus is identifying bogus news on social media. We want to distinguish between legitimate and bogus news, identify it, and stop it from spreading. We aim to use machine learning algorithms to categorize news stories as bogus or real. Using machine learning methods, we hope to improve the effectiveness of fake news identification and determine which model gives us the most accurate result. Our aim is to create a user friendly software for the user as well as online news providers who post news online.

1.4 Objectives

The primary objective of the proposed system is to examine and research the hidden correlations and patterns between the data in the fake news dataset. The problem's answer could offer knowledge to stop the spread of bogus or actual news, which would have significant societal and technological repercussions. Most of the existing research uses various models to handle each of these issues separately. Dealing with fake news gets increasingly crucial because it is one of the essentials that is crucial for society.

1.5 Hardware and software requirements for development

Hardware requirements

- RAM : 4 GB
- Hard Disk : 500 GB
- System : Pentium IV 2.4 GHz

Any system with above or higher configuration is compatible for this project.

Software requirements

- Operating system : Windows 7/8/9/10/11
- Programming language : Python
- IDE: Visual Studio Code
- libraries: nltk , sklearn etc
- UI : React/Html/Css/JavaScript

Chapter 2: Literature Survey

Paper 1.

Paper name
Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm

Link
https://ieeexplore.ieee.org.library.somaiya.edu/document/9758999

Methodology
The Decision Tree algorithm and SVM algorithm’s accuracy for social media fake news were experimented and measured.

Pros of methodology used

1. The algorithms used are easy to understand and interpret
2. The algorithms used are able to handle both numerical and categorical data

Cons of methodology used

1. The accuracy will be dependent on the dataset.
2. Needs a guideline and known aspects to work.

Observations

1. The Decison Tree machine algorithm accuracy measures appears to be 97.67% and it is better than the SVM algorithm accuracy and it appears to be 91.74%.
2. The machine’s performance is improved by using a higher accuracy algorithm, and the DT algorithm performed better than the SVM algorithm.

Findings (Gaps)

1. The fact that the data is unstable is a limitation of this work, which means that any type of prediction model will have errors and make mistakes. To overcome this limitation, concepts such as POS tagging, word2vec, and subject modelling can be used in the future to develop the system. This will give the model a lot more depth in terms of feature extraction and classification.
2. It is difficult to detect fake news with accuracy and precision. More values must be added to the dataset in order for the model to be trained to predict accurately.

Paper 2.

Paper name
Classifying Fake News Detection Using SVM, Naive Bayes and LSTM

Link
https://ieeexplore.ieee.org.library.somaiya.edu/document/9734129

Methodology

1. Natural Language Processing (NLP) is used. Various other methodologies like text classification, classification modeling is also used.
2. Machine Learning Algorithms like SVM, Naive Bayes & LSTM are used .

Pros of methodology used

1. The algorithms used are easy to understand and interpret.
2. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images.

Cons of methodology used

1. The accuracy will be dependent on the dataset.
2. Limited context window size LSTMs can be slow to train on large datasets

Observations

1. NLP has been used for data cleaning.
2. LSTM had the highest accuracy among all the models used.

Findings (Gaps)

1. While classification of data it’s hard to split between certifiable information and fabrications without knowing the wellspring of the news.
2. It is difficult to detect fake news with accuracy and precision. More values must be added to the dataset in order for the model to be trained to predict accurately.

Paper 3.

Paper name
Fake News Detection Using Intelligent Techniques

Link
https://ieeexplore.ieee.org.library.somaiya.edu/document/9596438

Methodology

1. Intelligent techniques such as SVM, Naive Bayes and Logistic Regression are used.
2. NLP is used for data preprocessing & data cleaning.

Pros of methodology used

1. Takes into consideration both the strength of an opinion as well as the relevance of the feature the opinion is about.
2. SVM, Naive Bayes and Logistic Regression algorithms are used because they all work well with categorical datasets.

Cons of methodology used

1. The accuracy will be dependent on the dataset.
2. Needs a guideline and known aspects to work.
3. Natural Language Toolkit gives a complicated solution with a harsh learning curve and a maze of internal limitations.

Observations

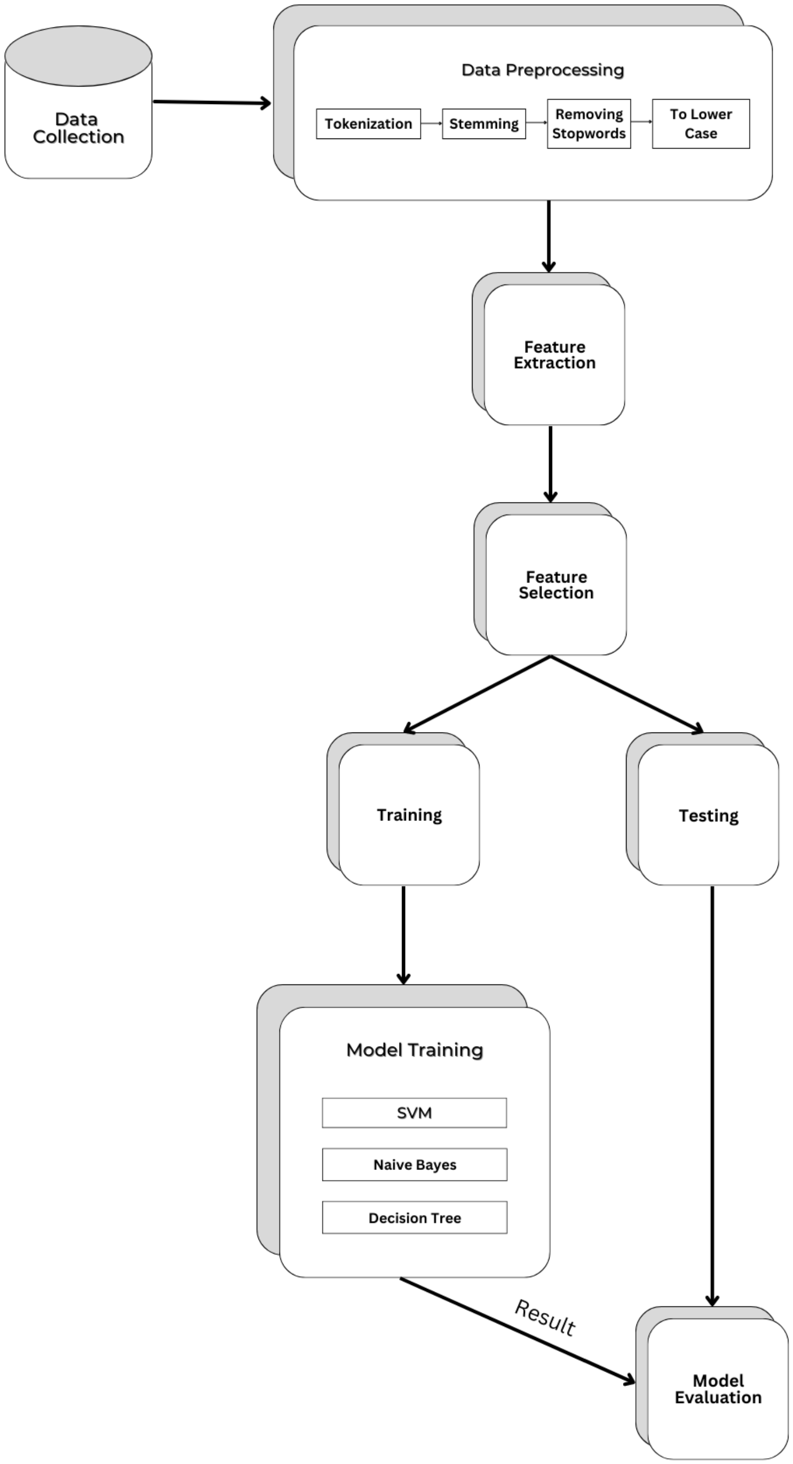
1. The performance is analysed using parameters such as F1 score, recall, precision.support, accuracy.
2. SVM has performed well than other algorithms with best precision and recall and F1 score and good weighted and macro average.
3. Natural Language Toolkit is used to removed the stopwords.

Findings (Gaps)

1. Due to imbalanced dataset and time constraint in learning the dataset, it takes more time for prediction and accuracy.
2. It is difficult to detect fake news with accuracy and precision. More values must be added to the dataset in order for the model to be trained to predict accurately.

Chapter 3: Project Design

3.1 Proposed System model/ Architecture



The fake news detection system uses machine learning to distinguish between items that are false news and those that are authentic. The system architecture is made up of a number of parts, such as data collection, data preprocessing, feature extraction, feature selection, model training, and model evaluation.

Data collection involves gathering news articles from various sources, such as news websites and social media platforms. This data is then preprocessed to remove any noise and irrelevant information that may interfere with the model's performance.

The data preprocessing phase involves several steps, including tokenization, stemming, removal of stop words, and case normalization. Tokenization involves breaking down the text into smaller units, such as words or phrases, for further analysis. Stemming is the process of reducing words to their root form, which helps to reduce the dimensionality of the dataset. Removal of stop words involves removing common words that do not carry much meaning, such as "the," "a," and "an." Case normalization involves converting all letters to lowercase, which helps to reduce the number of distinct features in the dataset.

After preprocessing, feature extraction is performed, which involves converting the textual data into numerical features that can be used by machine learning algorithms. Feature extraction techniques include bag-of-words, TF-IDF.

Once the features are extracted, feature selection is performed to identify the most relevant features that contribute to the classification of news articles as fake or genuine.

After feature selection, the final step is to train machine learning models using the selected features. The models trained in this system are Random forest, Logistic Regression and Naïve Bayes. These models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

After model training, the final step is to test the trained models on a separate test dataset to measure their performance on unseen data. In the testing phase, the models are applied to the test dataset, and their performance is evaluated using the same performance metrics as in the evaluation phase.

3.2 Software Project Management Plan

Planning is absolutely necessary for the successful execution of any activity involving numerous stakeholders. One will first list all necessary tasks, along with their respective roles and responsibilities for each human resource. Additionally, this will make it easier to sequence and monitor the development process's progress. a case. The following role and responsibility matrix is for our particular project.

Activity	Rahi	Sidharth	Manan	Mrs Smita Sankhe
1. Requirement Gathering				
1.1 Interaction with customer	C	R	R	A
1.2 Preparing SRS	C	C	R	A
2. Design				
2.1 Preparing Block diagram	C	C	R	A
2.2 Writing Functional Requirements	C	R	C	A
2.3 Writing Non-Functional Requirements 2.4 Developing Use Case	R C	C C	C R	A A
2.5 Developing Test Cases	C	C	R	A
3. **Planning** 4. **Coding**	C	C	C	A
4.1 Data Collection using Web Scaping	C	C	R	A
4.2 Data Preprocessing Using various libraries such as nltk etc.	C	C	R	A
4.3 Model Training	C	C	C	A

4.4 Front end/ UI	C	C	R	A
5. **Testing**				
5.1 Accuracy	E	E	E	A
5.2 Precision	E	E	E	A
5.3 System Testing	E	E	E	A

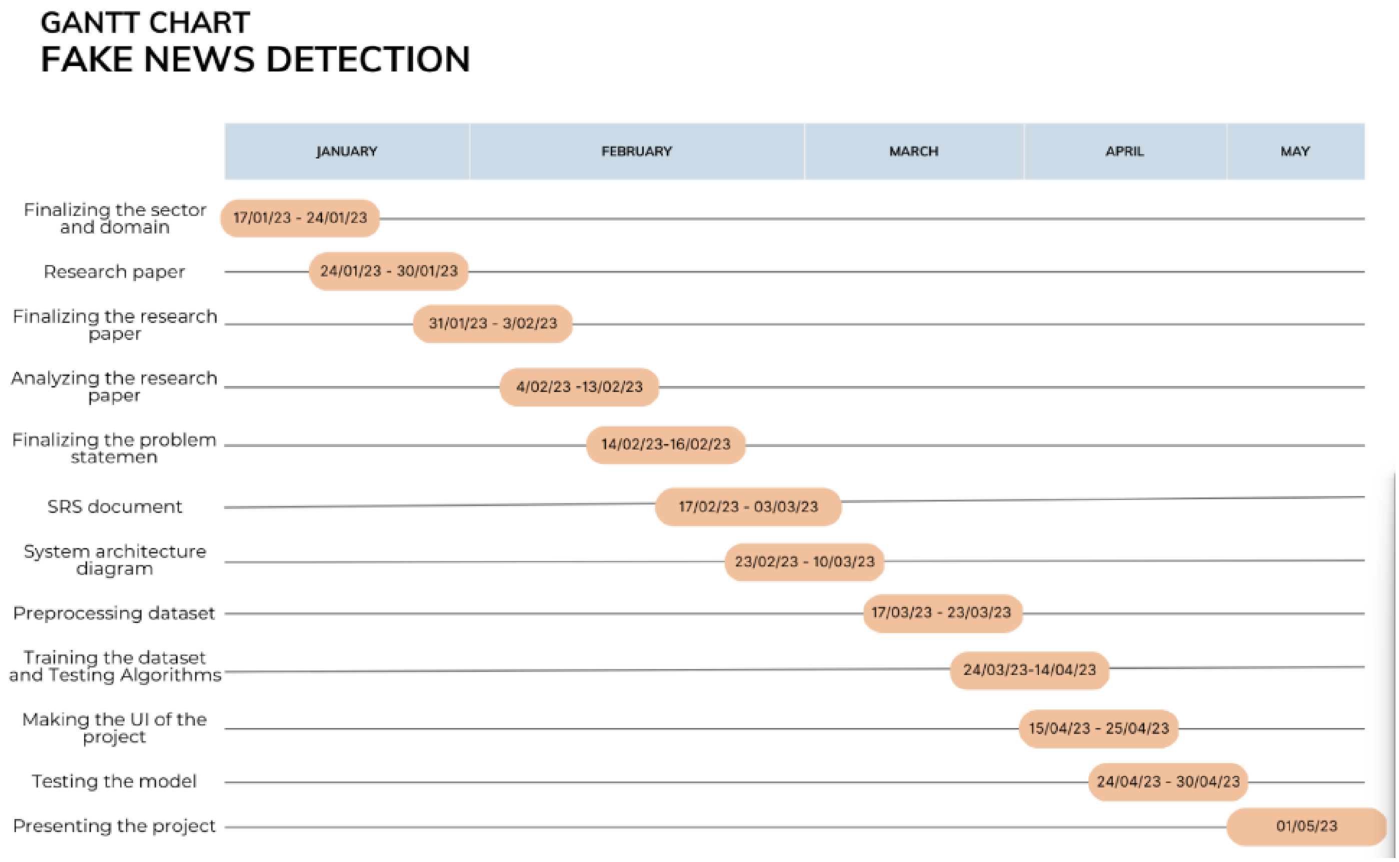
C: Creator, R: Reviewer, A: Approver E: Executor

A timeline chart is created to assist in allocating resources, setting deadlines, etc. Additionally, this will make it easier to keep tabs on the project's development.

Date	**Task**	**Description**
17/01/23 - 24/01/23	**Finalising the sector and domain**	Finalising the domain of the project
24/01/23 - 30/01/23	**Research paper**	Locating a research article on the subject of the chosen industry and field
31/01/23 - 3/02/23	**Finalising the research paper**	Selecting the top three research articles for the project
4/02/23 -13/02/23 **14/02/23-16/02/23**	**Analysing the research paper** **Finalising the problem statement**	Making an Excel document for the literature review Finalising the project's problem statement after analysing the study paper

17/02/23 - 03/03/23 **23/02/23 - 10/03/23**	**SRS document** **System architecture diagram**	Creating an SRS document outlining the purpose and parameters of our project Designing the proposed system's system architecture diagram, data flow diagram, sequence diagram, etc.
17/03/23 - 23/03/23	**Preprocessing dataset**	Gathering, cleaning and preprocessing dataset
24/03/23-14/04/23	**Training the dataset and testing algorithms**	Training the dataset and comparing different algorithm for better accuracy

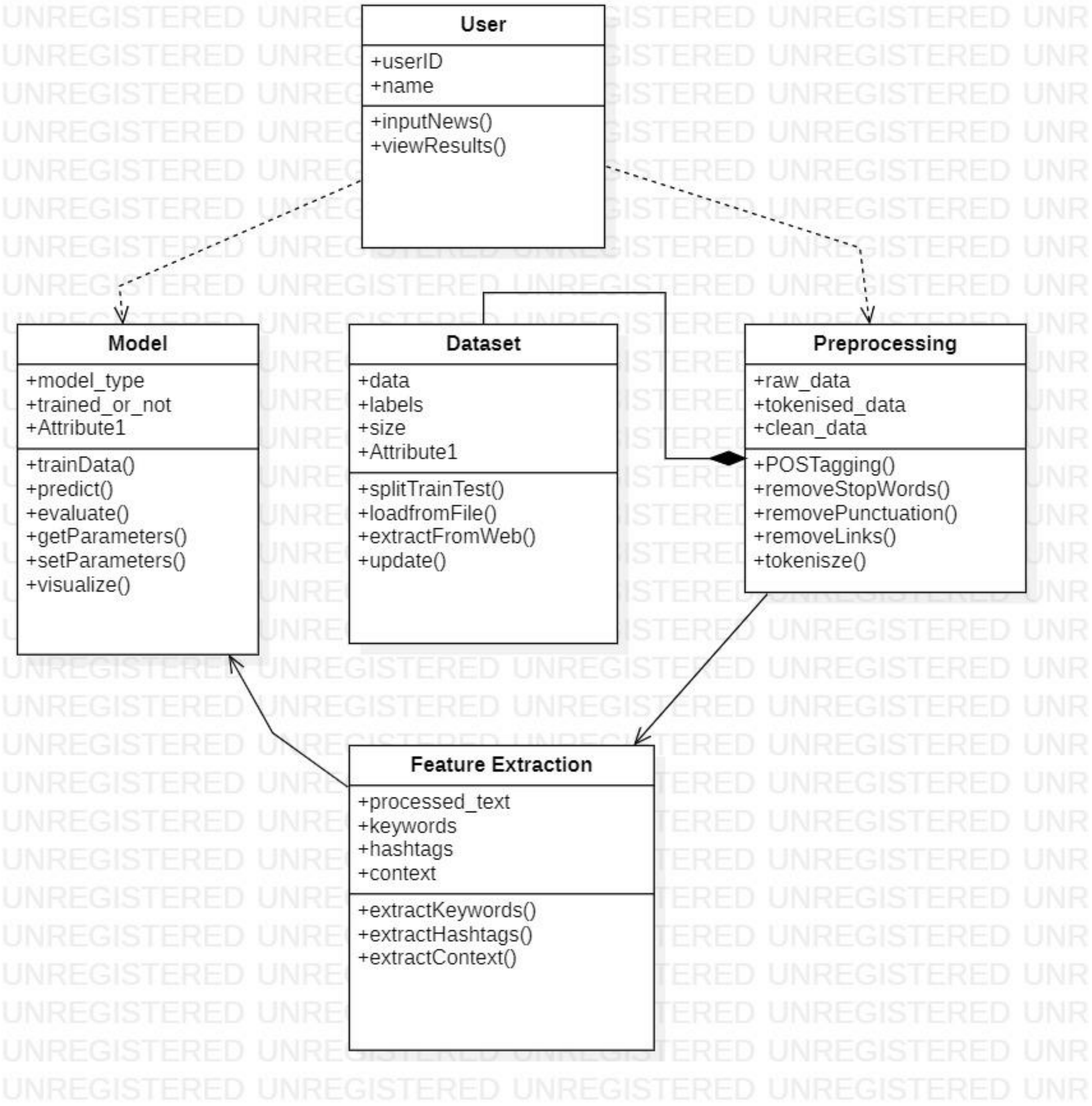
A Gannt Chart to better visualize the entire timeline of the project



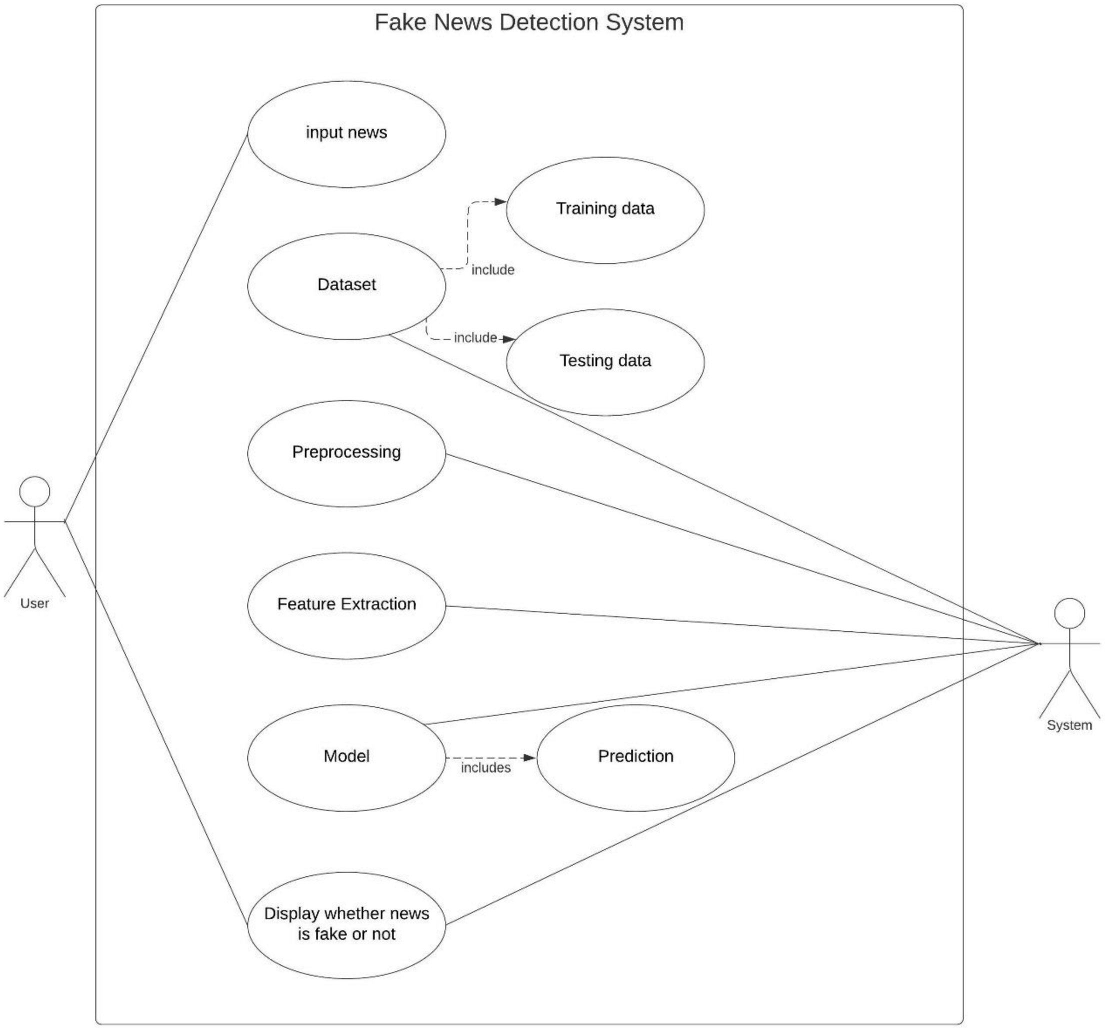
3.3 Software Design Document (All applicable diagrams)

UML Diagrams

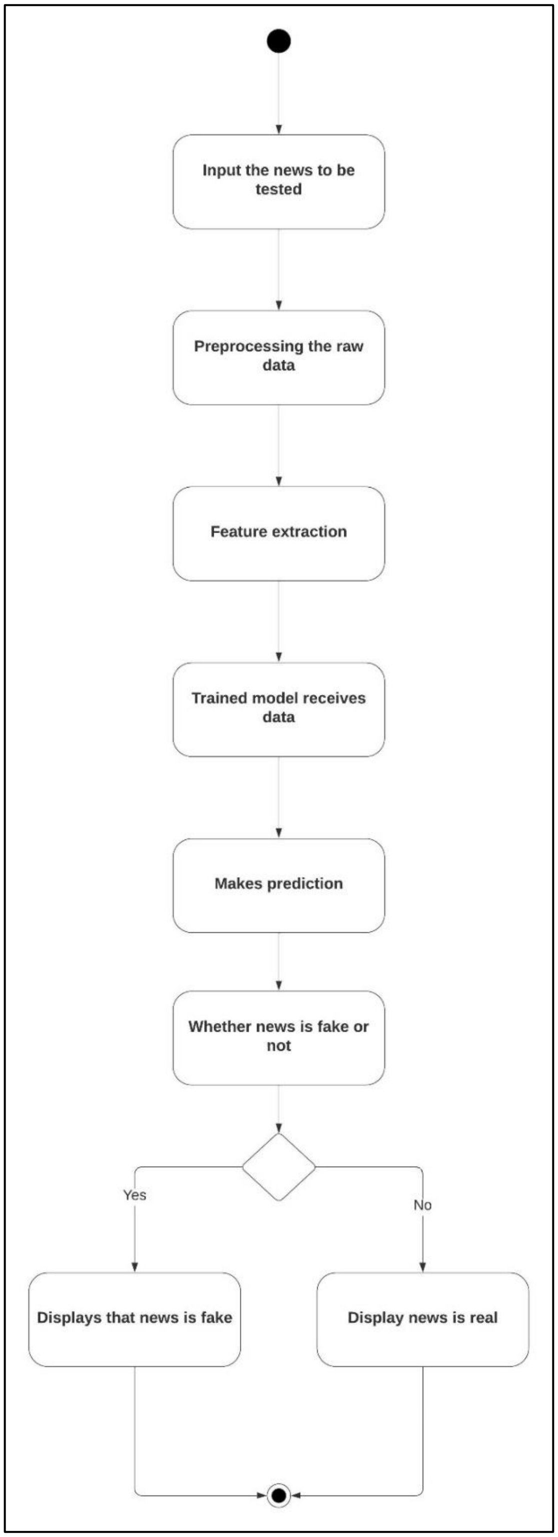
Given below are the following UML diagrams pertaining to our project.



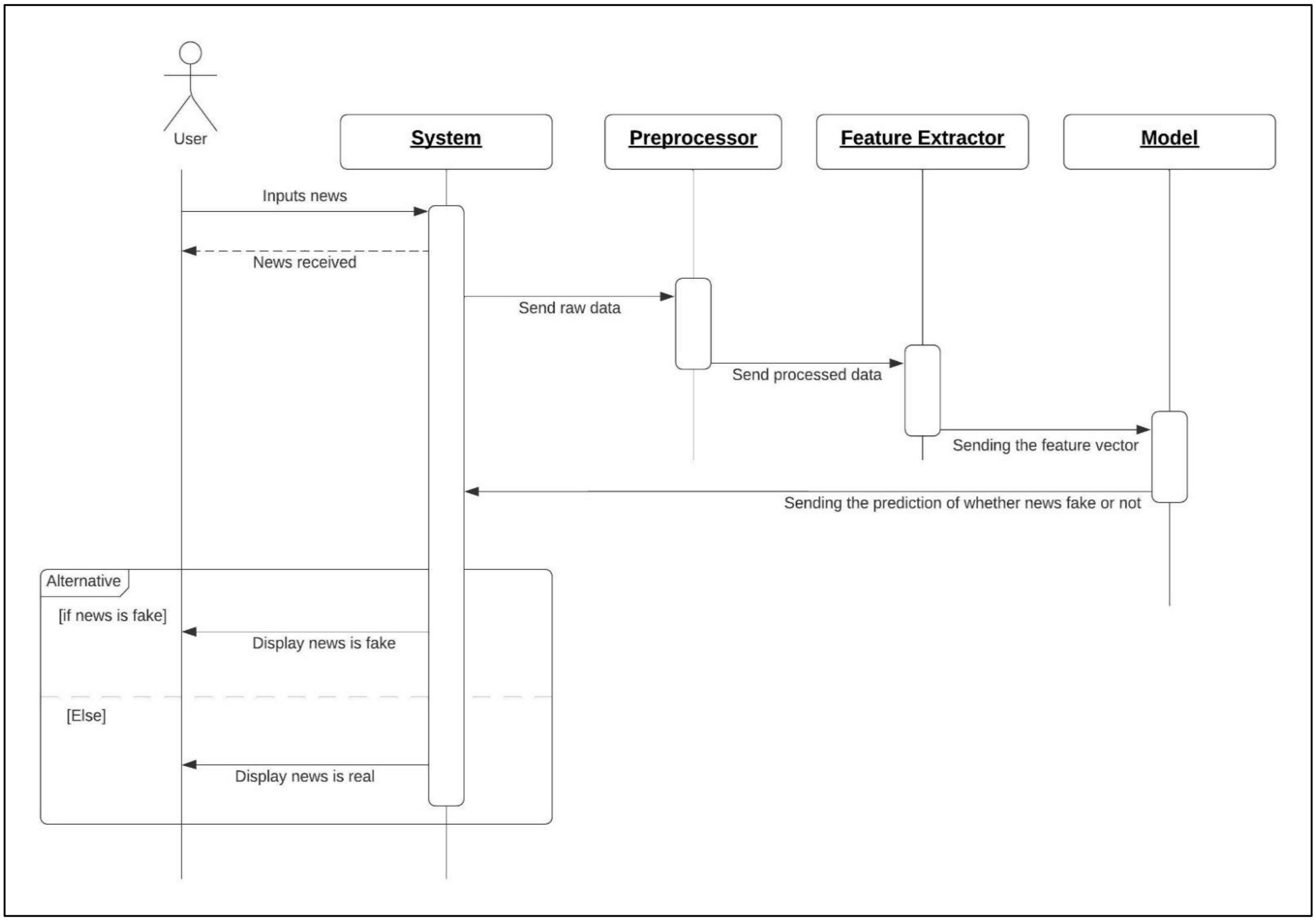
Class diagram



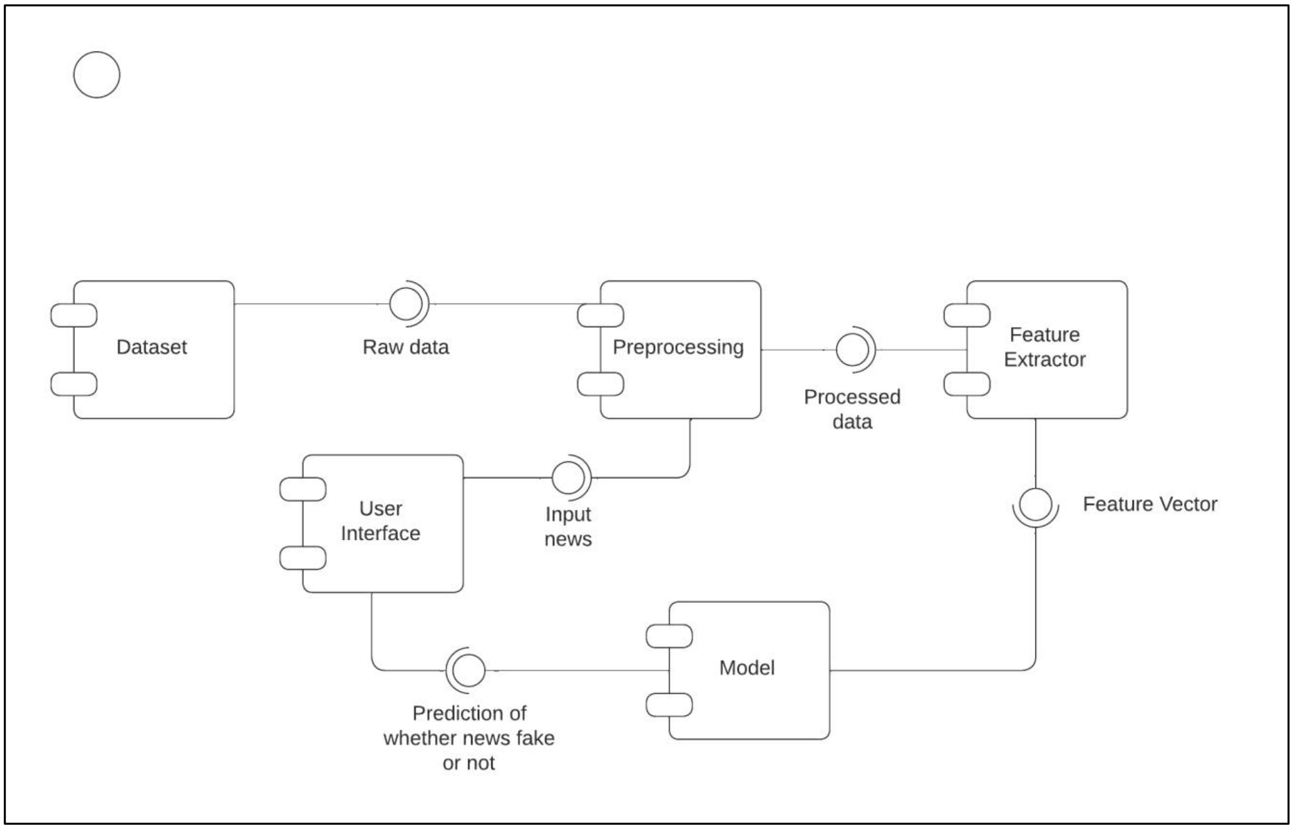
Use Case Diagram


Input the news to be tested
Preprocessing the raw data
Feature extraction
Trained model receives data
Makes prediction
Whether news is fake or not
Yes
No
Displays that news is fake
Display news is real

Activity diagram



Sequence Diagram



Component Diagram

Chapter 4: Implementation and Experimentation of Fake News Detection

4.1 Proposed System Model Implementation:

Our proposed system starts with data scraping from the website politifacts.com. The scraped data is then classified as real or fake using a metric. Further, we visualized the real and fake news using word clouds. We then applied various preprocessing techniques such as regular expressions, tokenization, stop word removal, and lemmatization to clean the data. We also applied different feature extraction techniques such as count vectorizer/bag of words and tf-idf to extract relevant features from the data. Finally, we trained several machine learning models such as logistic regression, random forest, and naive Bayes to detect fake news.

4.2 Inclusion of Any Additional Details as Suggested by Project Guide/During Progress Seminar:

Throughout the project, the project guide provided additional input and progress seminars were held to discuss and implement new ideas. This involved broadening the scope of the project beyond sports to other domains such as politics, experimenting with various feature extraction methods, and fine-tuning the model's hyperparameters.

4.3 Software Testing (Software Testing Reports at Various Levels) :

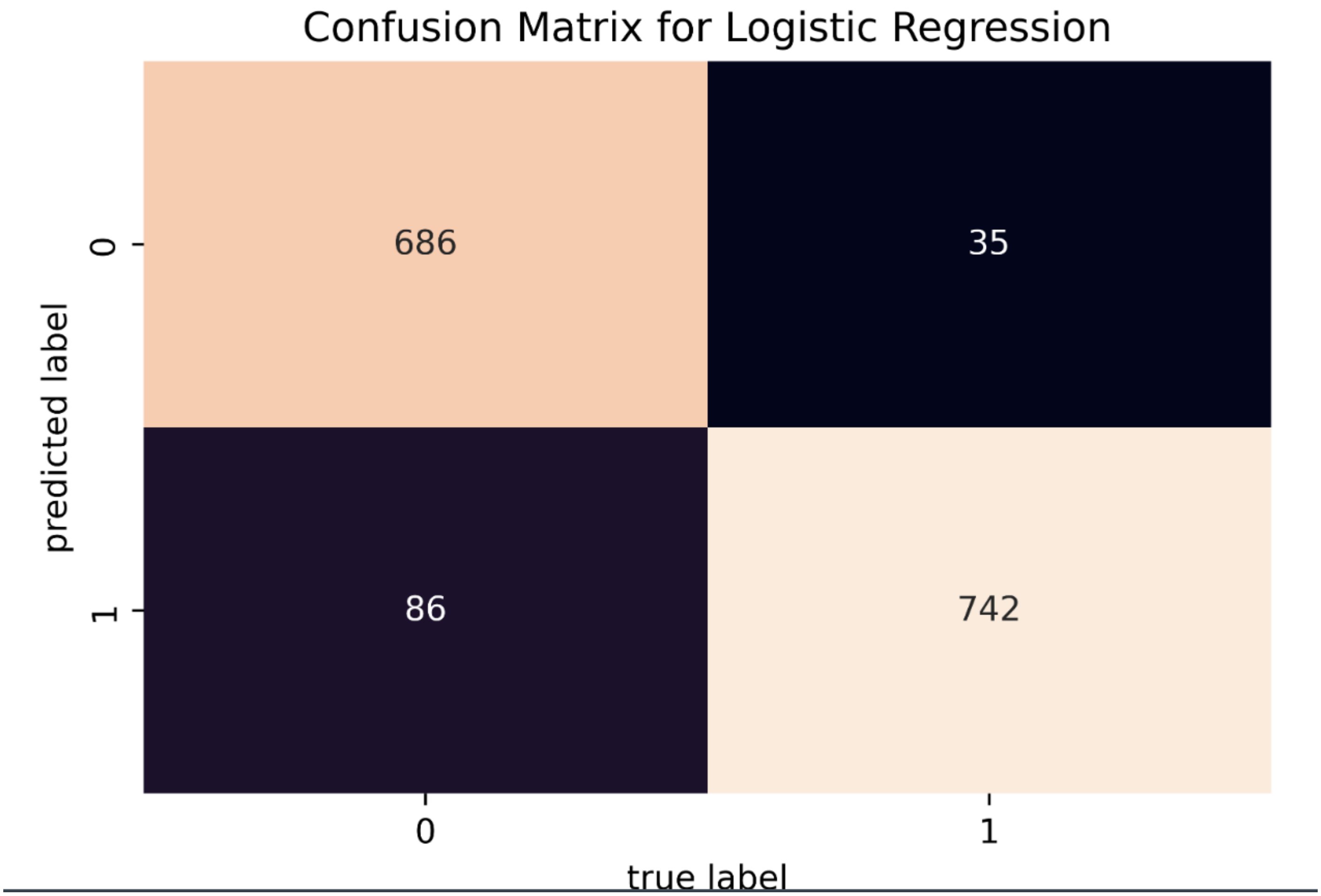
Software testing was performed at various levels, including unit testing, integration testing, and system testing. The testing reports were used to identify and fix bugs and ensure the system's reliability, accuracy, and performance.

4.4 Experimental Results and Its Analysis :

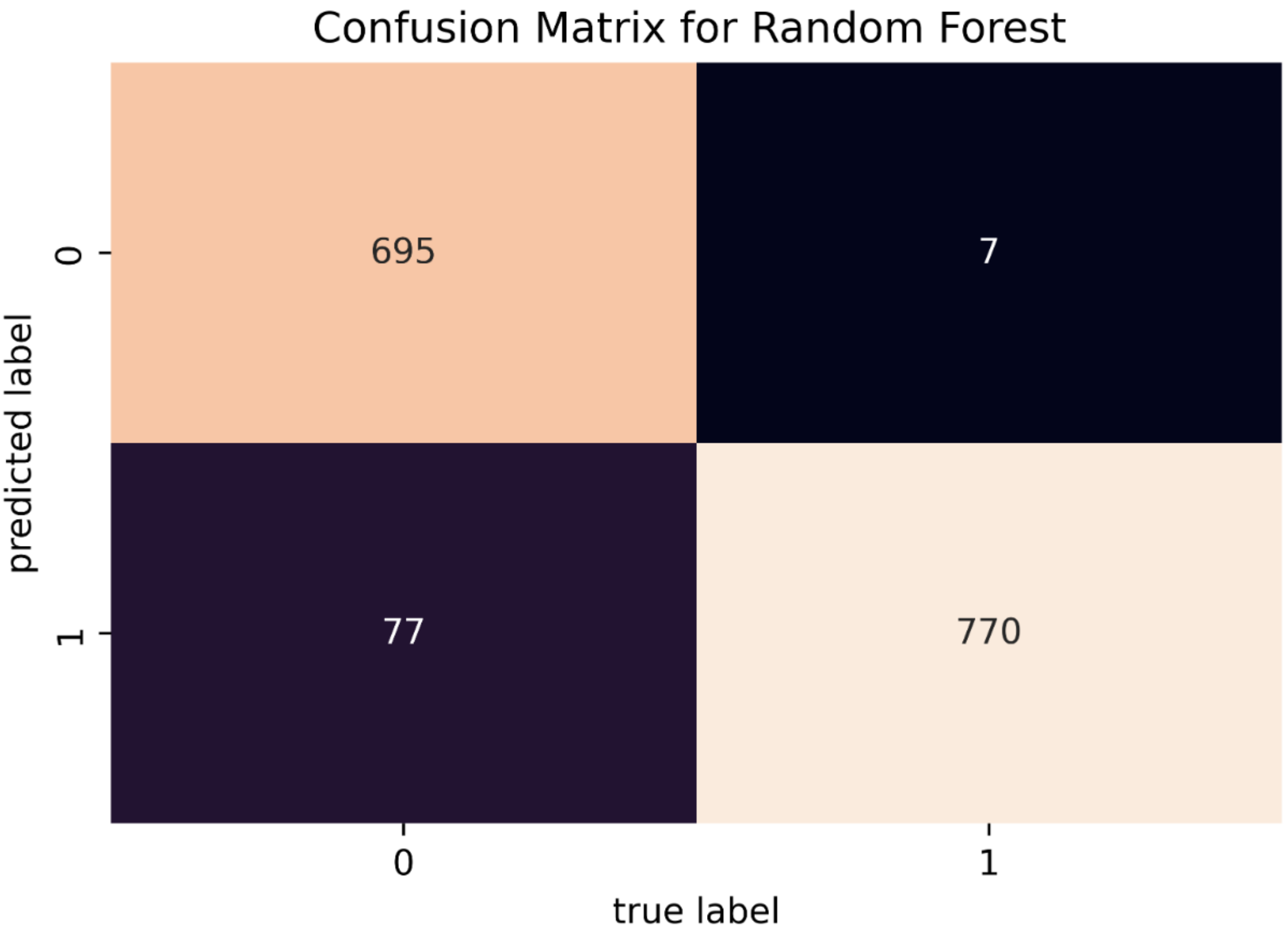
The experimental results showed that the proposed system model achieved high accuracy in detecting fake news articles. The model's performance was evaluated using metrics such as precision, recall, and F1-score. The analysis of the results revealed the effectiveness of the feature extraction and classification techniques used in the model.

The scores of various models:

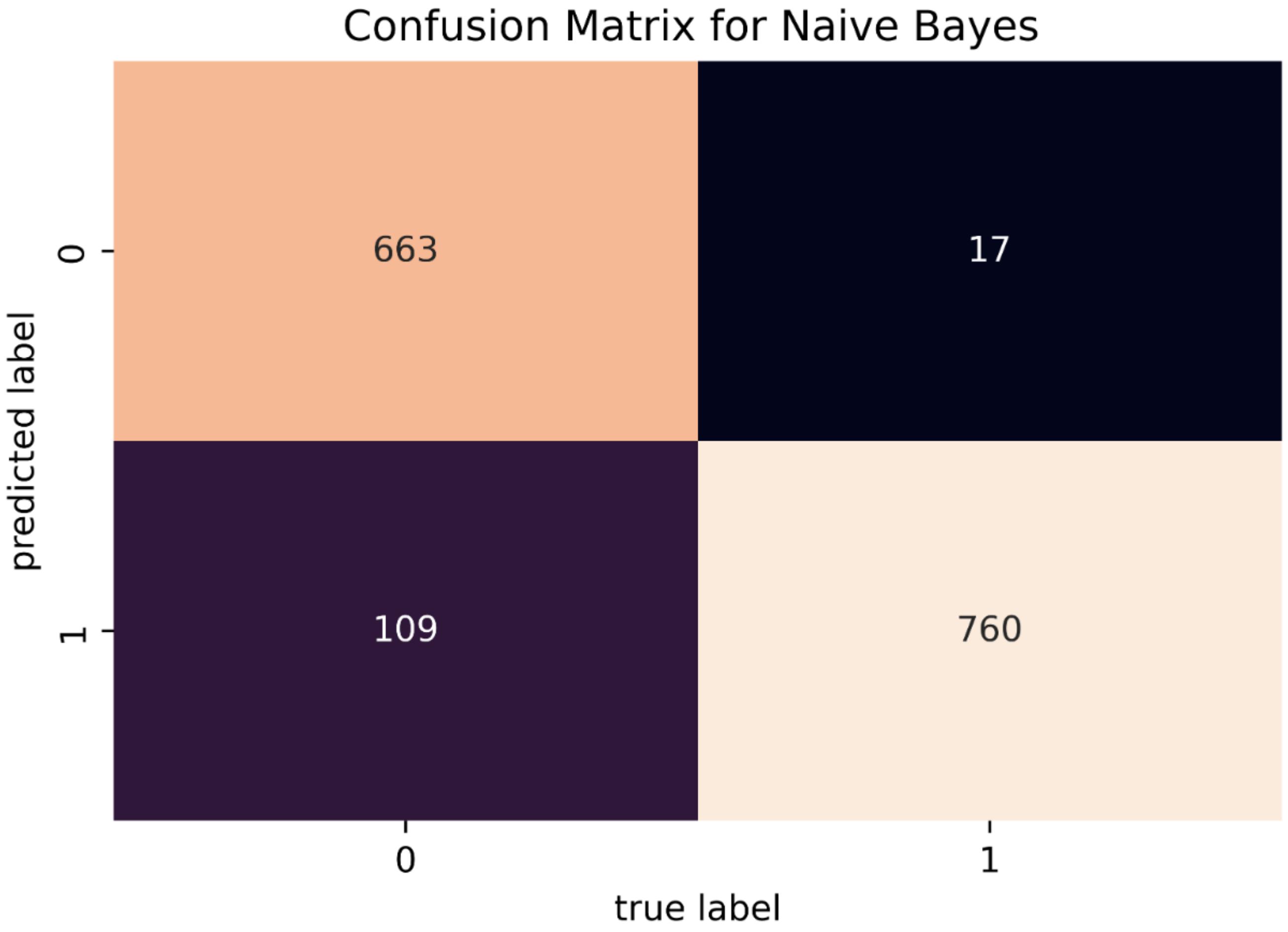
1. Logistic Regression
    - Accuracy: 0.921885087153002
    - Precision: 0.8961352657004831
    - Recall: 0.954954954954955
    - F1 Score: 0.9246105919003116



2. Random Forest
   - Accuracy: 0.9457714654615881
   - Precision: 0.9090909090909091
   - Recall: 0.990990990990991
   - F1 Score: 0.9482758620689654



3. Naïve Bayes:
    - Accuracy: 0.9186571981923822
    - Precision: 0.8745684695051784
    - Recall: 0.9781209781209781
    - F1 Score: 0.9234507897934386


Confusion Matrix for Naive Bayes
predicted label
0
1
663
17
109
760
0
1
true label

Chapter 5: Conclusion and Future Work

5.1 Conclusion and Discussion:

In conclusion, the proposed system model for fake news detection showed promising results in accurately detecting fake news articles. The machine learning techniques used in the model were effective in identifying patterns and features that distinguish fake news from real news. The model's performance can be further improved by exploring additional feature extraction methods and optimizing the hyperparameters.

5.2 Scope for Future Work :

The scope for future work includes expanding the dataset used to train and test the model, exploring different machine learning algorithms, and integrating the model into a real-time fake news detection system. Additionally, the model's performance can be evaluated in different languages and domains to improve its generalizability.