A

Project Report On

Fake News Detection

Submitted in partial fulfillment of the requirement for the VI semester

**Bachelor of Computer Science**

Submitted By

**Shailesh Joshi**

Under the Guidance of **Mrs. Shilpa Jain Assistant Professor Department of CSE**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING GRAPHIC ERA HILL UNIVERSITY, BHIMTAL CAMPUS SATTAL ROAD, P.O. BHOWALI,**

**DISTRICT- NAINITAL-263132 2023- 2024**

# STUDENT’S DECLARATION

We, **Shailesh Joshi** here by declare the work, which is being presented in the project, entitled “**Fake News Detection**” in partial fulfillment of the requirement for the award of the degree **B.Tech** in the session **2023-2024**, is an authentic record of my own work carried out under the supervision of “**Mrs. Shilpa Jain ”, Assistant Professor, Department of CSE, Graphic Era Hill University, Bhimtal.**

The matter embodied in this project has not been submitted by us for the award of any other degree.

**Date: 12 / 06 / 2024**

**Shailesh Joshi**

# CERTIFICATE

The project report entitled **“Fake News Detection”** being submitted by

**Shailesh Joshi** to Graphic Era Hill University Bhimtal Campus for the award

of bonafide work carried out by them. They have worked under my guidance and supervision and fulfilled the requirement for the submission of report.

**(Mrs. Shilpa Jain) (Dr. Ankur Bisht)**

**Project Guide (HOD, CSE Dept.)**

# ACKNOWLEDGEMENT

We take immense pleasure in thanking Honorable **“Mrs. Shilpa Jain”** (**Assistant Professor, CSE, GEHU Bhimtal Campus**) to permit me and carry out this project work with his excellent and optimistic supervision. This has all been possible due to his novel inspiration, able guidance and useful suggestions that helped me to develop as a creative researcher and complete the research work, in time.

Words are inadequate in offering my thanks to GOD for providing me everything that we need. We again want to extend thanks to our President **“Prof. (Dr.) Kamal Ghanshala”** for providing us all infrastructure and facilities to work in need without which this work could not be possible.

Many thanks to Professor **“Dr. Manoj Chandra Lohani”** (Director Gehu Bhimtal), other faculties for their insightful comments, constructive suggestions, valuable advice, and time in reviewing this thesis.

Finally, yet importantly, we would like to express my heartiest thanks to our beloved parents,

for their moral support, affection and blessings. We would also like to pay our sincere thanks to all our friends and well-wishers for their help and wishes for the successful completion of this research.

**Shailesh Joshi**

**ABSTRACT**

In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by these online fake news easily, which has brought about tremendous effects on the offline society already. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely. This paper aims at investigating the principles, methodologies and algorithms for detecting fake news articles, creators and subjects from online social networks and evaluating the corresponding performance. Information preciseness on Internet, especially on social media, is an increasingly important concern, but web-scale data hampers, ability to identify, evaluate and correct such data, or so called "fake news," present in these platforms. In this paper, we propose a method for "fake news" detection and ways to apply it on Facebook, one of the most popular online social media platforms. This method uses Naive Bayes classification model to predict whether a post on Facebook will be labeled as real or fake. The results may be improved by applying several techniques that are discussed in the paper. Received results suggest, that fake news detection problem can be addressed with machine learning methods

**TABLE OF CONTENTS**

Declaration II

[Certificate III](#_TOC_250002)

[Acknowledgement… IV](#_TOC_250001)

Abstract…………………………………………………………………………………V

[Table of Contents… V](#_TOC_250000)I-VII

CHAPTER 1: INTRODUCTION… VIII

* 1. Objective ix
  2. Problem Statement… x
  3. Methodology… xi-xii
  4. Project Organization… xiii

CHAPTER 2: S/W AND H/W REQUIREMENTS (UP TO FULLEST EXTENT)

… XIV

* 1. S/W and H/W requirements (up to fullest extent) xiv
  2. Resources and Technology used… xv

CHAPTER 3: FORMULAS AND WORKING… XVI-XVIII

CHAPTER 4: DFD (AT LEAST UP TO 2 LEVEL) XIX

* 1. DFD (At least up to 2 level) xix

CHAPTER 5: ER DIAGRAM XXI

* 1. ER Diagram… xxi

CHAPTER 5: ACTIVITY DIAGRAM XXII

CHAPTER 6: IMAGES OF CODE\_SNIPPET XXIII-XXIV

CHAPTER 7: CONCLUSION XXVII

REFERRENCE XXVIII

# INTRODUCTION

These days‟ fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is “fake news “ but lately blathering social media‟s discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints. The importance of disinformation within American political discourse was the subject of weighty attention , particularly following the American president election . The term 'fake news' became common parlance for the issue, particularly to describe factually incorrect and misleading articles published mostly for the purpose of making money through page views. In this paper,it is seeked to produce a model that can accurately predict the likelihood that a given article is fake news.Facebook has been at the epicenter of much critique following media attention. They have already implemented a feature to flag fake news on the site when a user sees‟s it ; they have also said publicly they are working on to to distinguish these articles in an automated way. Certainly, it is not an easy task. A given algorithm must be politically unbiased – since fake news exists on both ends of the spectrum – and also give equal balance to legitimate news sources on either end of the spectrum. In addition, the question of legitimacy is a difficult one.However, in order to solve this problem, it is necessary to have an understanding on what Fake News.

# OBJECTIVE

The objective of this project is to examine the problems and possible significances related with the spread of fake news. We will be working on different fake news data set in which we will apply different machine learning algorithms to train the data and test it to find which news is the real news or which one is the fake news. As the fake news is a problem that is heavily affecting society and our perception of not only the media but also facts and opinions themselves. By using the artificial intelligence and the machine learning, the problem can be solved as we will be able to mine the patterns from the data to maximize well defined objectives. So, our focus is to find which machine learning algorithm is best suitable for what kind of text dataset. Also, which dataset is better for finding the accuracies as the accuracies directly depends on the type of data and the amount of data. The more the data, more are your chances of getting correct accuracy as you can test and train more data to find out your results

.

# PROBLEM STATEMENT



The major problem in detecting fake news is the lack of a massive dataset and a labeled benchmark dataset with ground-truth labels. For example, some of the datasets are constructed only with political statements like PolitiFact, LIAR, Weibo, etc.

# METHODOLOGY

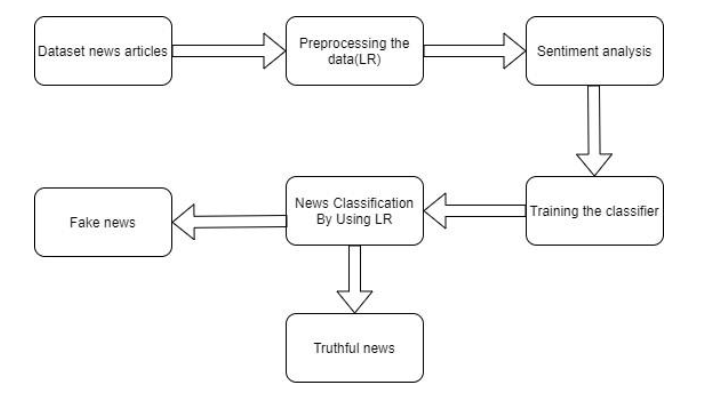
# Existing System:

# There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature. Conroy, Rubin, and Chen outlines several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

**Proposed System:**

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

**System Architechture:**



# PROJECT ORGANISATION



This Project is created by the team of 1 People:

* The Data Processing and applying the Machine Learning Algorithm is Done by Shailesh Joshi
* And the Front end Part creating the HTML webpage is also done by Shailesh Joshi

# S/W AND H/W REQUIREMENTS

# Hardware Requirements

* A PC with Windows/Linux OS
* Processor with 1.7-2.4gHz speed
* Minimum of 8gb RAM
* 2gb Graphic card

**Software Specification**

* Text Editor (VS-code/WebStorm)
* Python libraries
* Coding Language-Python

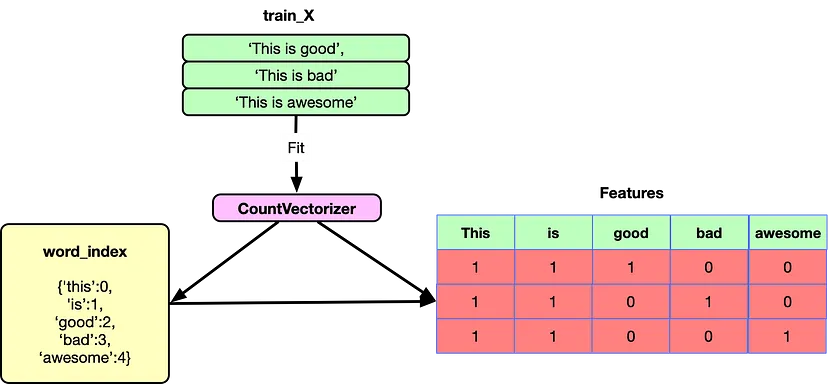
# RESOURCES AND TECHNOLOGY USED

**For making simple & attractive front-end and other tools for backend:**

Our project will be a web application and it should be responsive for each accessible device. Hence we are using these Technologies for making attractive front-end.

* **HTML 5:** Since it is a web application that has to be displayed in a browser and we know that browser display everything in the form of markup. HTML is a markup language that provides the structure of a website so that web browsers know what to show. That’s why we are using Hyper Text Markup Language (HTML);
* **CSS :** Cascading Style Sheets, commonly known as CSS, is an integral part of the modern web development process. It is a highly effective HTML tool that provides easy control over layout and presentation of website pages by separating content from design.
* **Pandas:** Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Pandas provides in-memory 2d table object called Data frame.
* **Flask**: It is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications.

**WORKING**

****

The term TF-idf model is used to preprocess the text by converting it into a count, which keeps a count of the total occurrences of most frequently used words.

For Example:

we construct a vector, which would tell us whether a word in each sentence is a frequent word or not. If a word in a sentence is a frequent word, we set it as 1, else we set it as 0.

This can be implemented with the help of following code:

X = []

for data in dataset:

vector = []

for word in freq\_words:

if word in nltk.word\_tokenize(data):

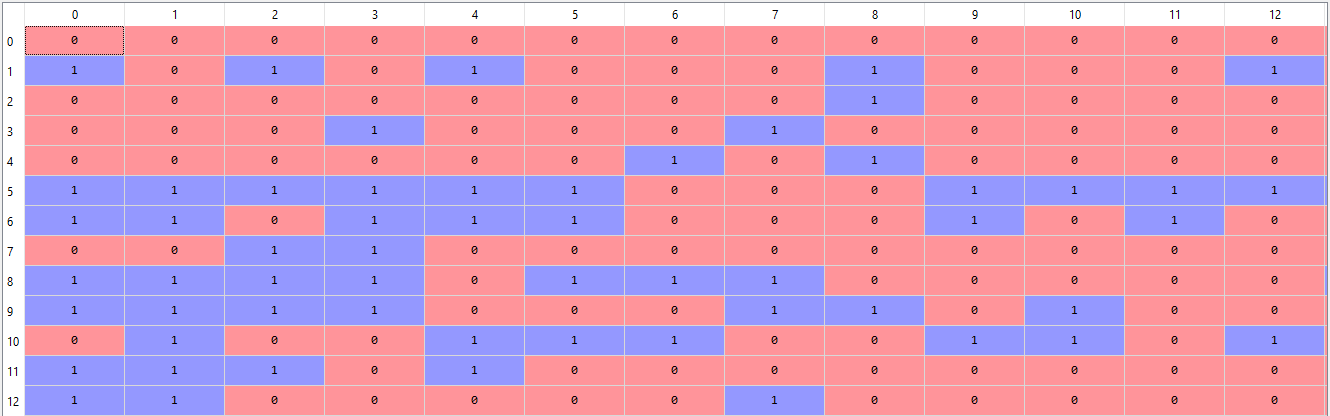
vector.append(1)

else:

vector.append(0)

X.append(vector)

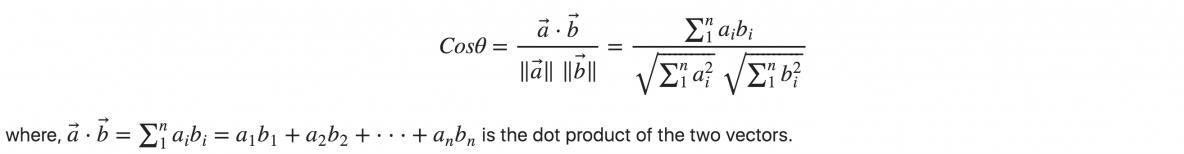
X = np.asarray(X)



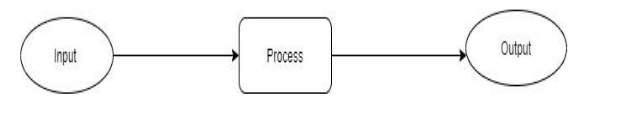
**FORMULAS USED**

**Cosine Similarity: Cosine similarity** is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

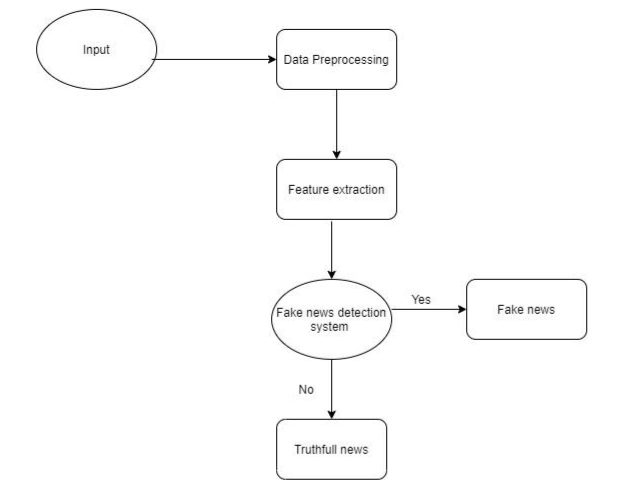
Formula:



**DFD**

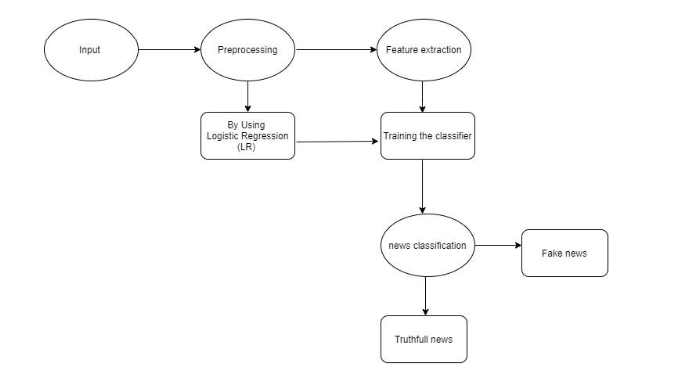
****

**Fig:** DFD 0

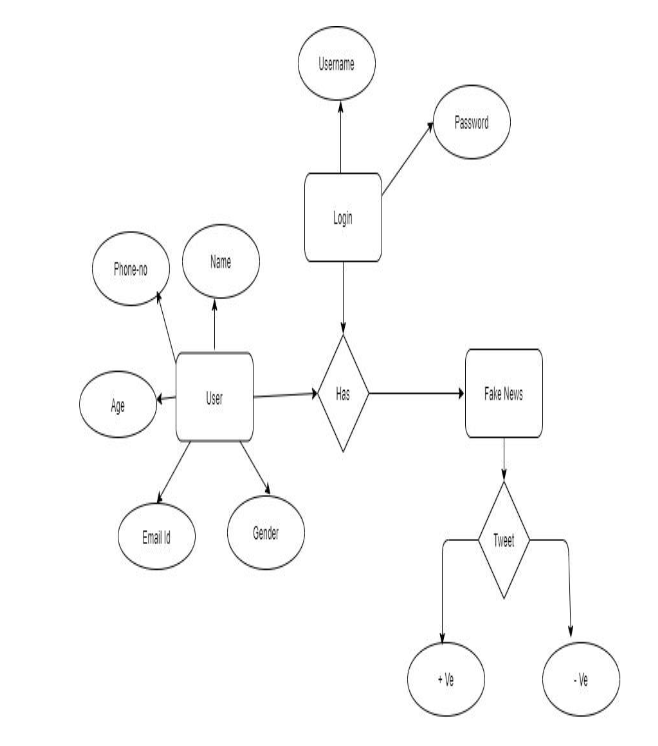
****

**Fig:** DFD 1



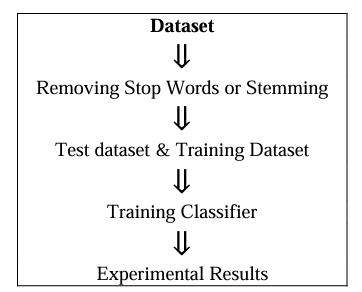
****

**ER-Diagram**

 ****

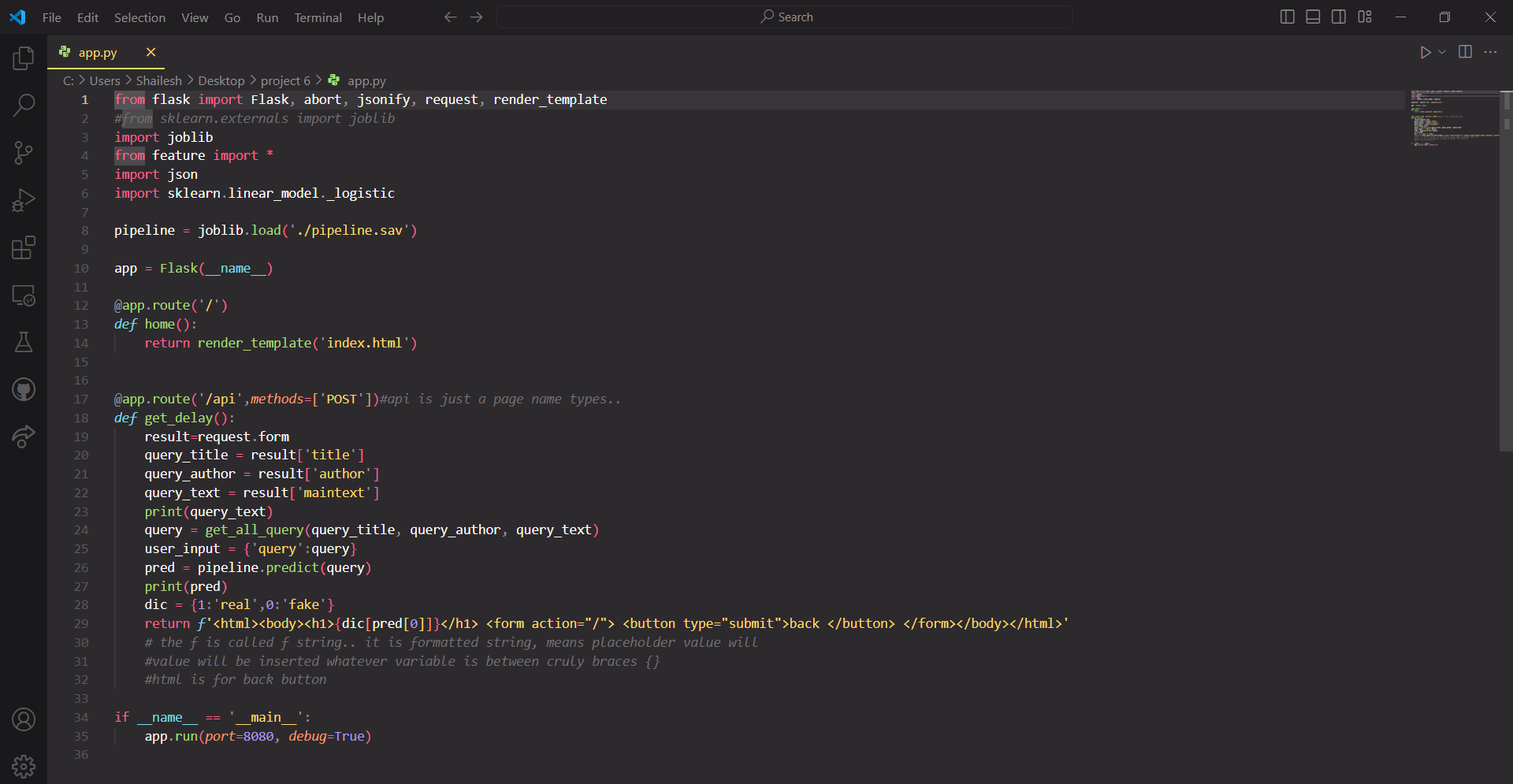
**ACTIVITY DIAGRAM**

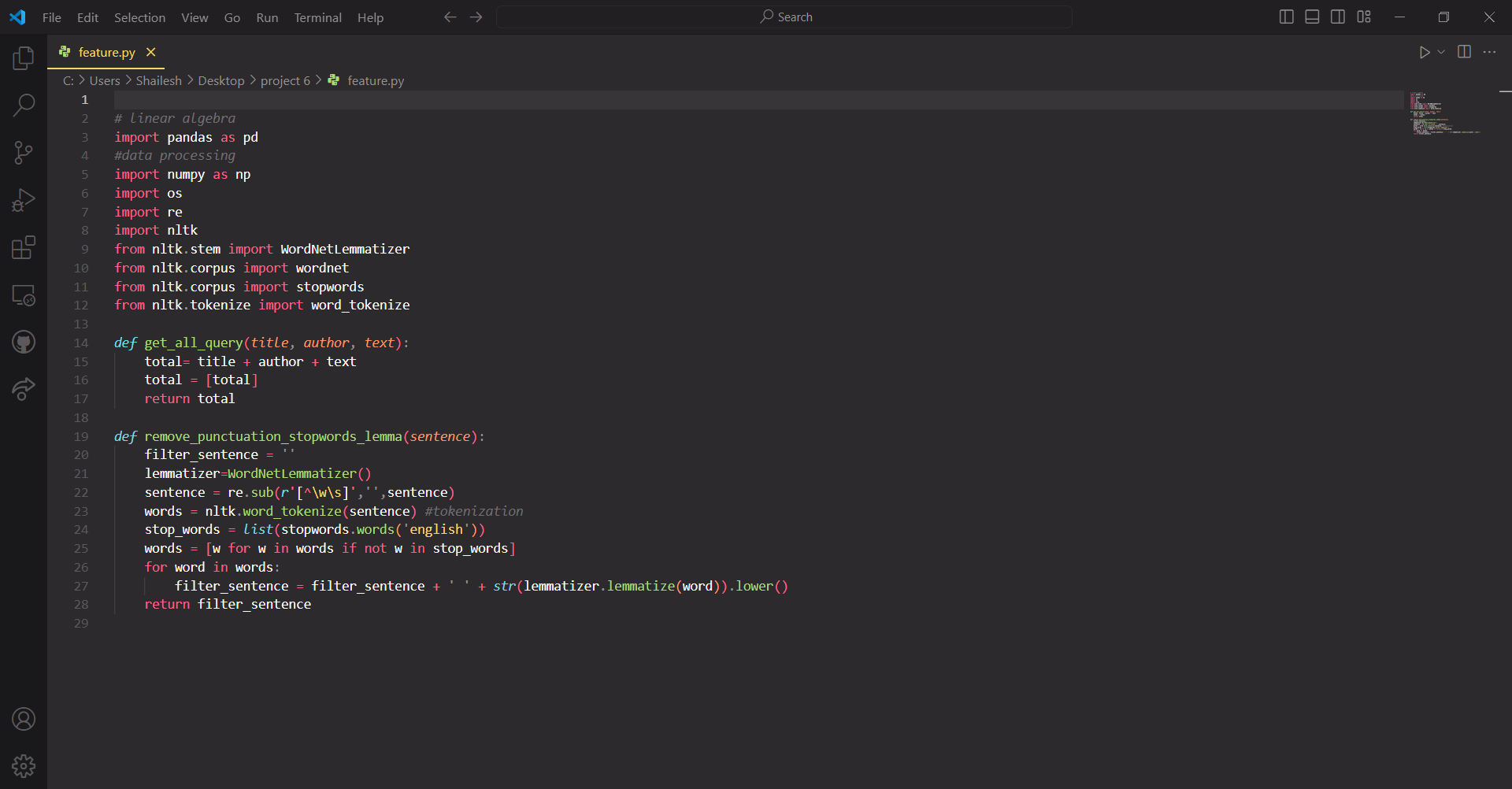


****

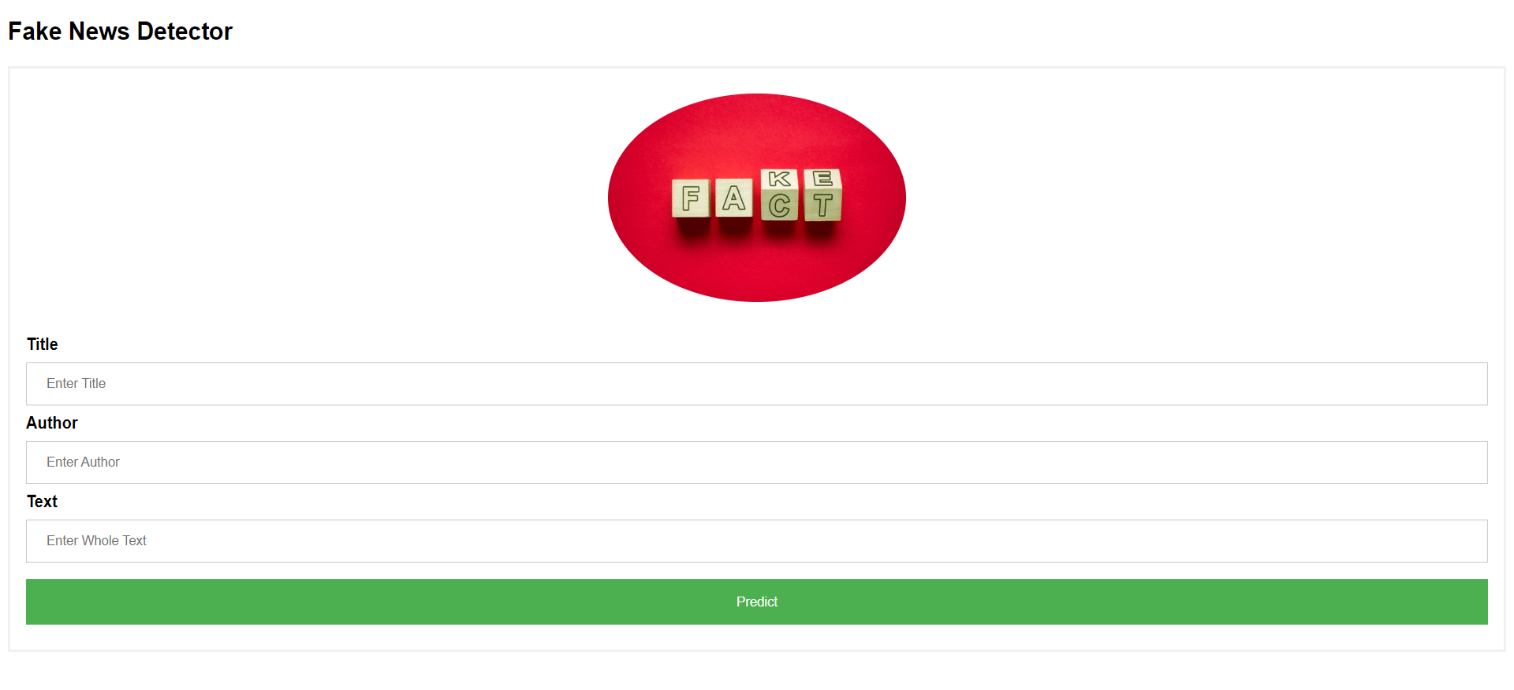
**IMAGES OF CODE**

**Front end (display page):**

****

****

**Website screenshot:**

****

********

# CONCLUSION

Due to increasing use of internet, it is now easy to spread fake news. A huge number of persons are regularly connected with internet and social media platforms. There is no any restriction while posting any news on these platforms. So some of the people takes the advantage of these platforms and start spreading fake news against the individuals or organizations. This can destroy the repute of an individual or can affect a business. Through fake news, the opinions of the people can also be changed for a political party. There is a need for a way to detect these fake news. Machine learning classifiers are using for different purposes and these can also be used for detecting the fake news. The classifiers are first trained with a data set called training data set. After that, these classifiers can automatically detect fake news. In this systematic literature review, the supervised machine learning classifiers are discussed that requires the labeled data for training. Labeled data is not easily available that can be used for training the classifiers for detecting the fake news. In future a research can be on the use of the unsupervised machine learning classifiers for the detection of fake news.

# REFERENCES

1. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” SIGKDD Explor. Newsl., vol. 19, no. 1, sep 2017.
2. The dataset contains two types of articles fake and real News. This dataset was collected from realworld sources; the truthful articles were obtained by crawling articles from Reuters.com (News website). As for the fake news articles, they were collected from different sources. The fake news articles were collected from unreliable websites that were flagged by Politifact (a fact-checking organization in the USA) and Wikipedia. The dataset contains different types of articles on different topics, however, the majority of articles focus on political and World news topics.
3. The dataset consists of two CSV files. The first file named “True.csv” contains more than 12,600 articles from reuter.com. The second file named “Fake.csv” contains more than 12,600 articles from different fake news outlet resources. Each article contains the following information: article title, text, type and the date the article was published on. To match the fake news data collected for kaggle.com, we focused mostly on collecting articles from 2016 to 2017. The data collected were cleaned and processed, however, the punctuations and mistakes that existed in the fake news were kept in the text.
4. This Dataset Is Provided by Kaggle and with the help of that only, we are able to Identify,Test and Train our Model