1. INTRODUCTION

1.1 ABOUT THE PROJECT

Fake news is described as a story that is made up with an intention to misdirect or to delude the reader. Fake news is one of the most serious problems in the modern age of the internet and social media. Since the internet is pervasive, everybody depends on different online resources for news. Fake news impacts have expanded exponentially in the past and something must be done to keep this from proceeding later in the future.

The key goal of this system is to identify fake news, which is a big issue in today's world. It is necessary to develop a model that can distinguish between "true" and "fake" news. So, Data mining and machine learning techniques can provide an efficient approach where data associated with both real and fake news can be used for classifying the authenticity of given news article.

1.2 OBJECTIVE

Our main objective is to make use of Machine Learning algorithms to create a model that can identify genuinity of given news article by training the model on fake and real news data that we have gathered from various sources.

This data contains all the details about various spatial features that can be helpful in classifying the news article. Moreover, Machine learning also provides results with high accuracy compared to the existing systems.

2. SYSTEM ANALYSIS

2.1 EXISTING SYSTEM

The existing models are based on Deep Learning Techniques, limitation of such approaches is that they are time consuming and costly. Neural Networks will also require much more data than an everyday person might have on hand to actually be effective. It gives less accuracy because of incomplete data. They also contain outdated input which is a problematic issue when it comes to designing a model.

2.2 PROPOSED SYSTEM

Our proposed system makes use of Machine Learning algorithms for classifing news articles. The dataset we have included consists of various news articles both real and fake news. The workflow reads the data and trains a regression model based on various attributes. The model is then used to predict the correctness of the article. The proposed system focuses on achieving high accuracy which helps people to follow real news.It uses Random Forest Algorithm to classify the news articles.

2.2.1 Details

Initially, we import the dataset that we have obtained from the Internet which has all the latest news articles. Then, we pre-process the dataset by removing all the unnecessary columns and also draft few plots to understand the trends,categories of news. We also transform the columns into our required format so that training the model on these would become easier.

Later, we split the dataset into training and testing parts and apply Random Forest regression algorithm which then fits the data into the model and predicts the output (i.e. classifying the authenticity of given news article) for the input conditions we give.

2.2.2 Impact on Environment

Fake news impacts have expanded exponentially. Fake news is one of the most serious problems in the modern age of the internet and social media This can create confusion and misunderstanding about important social, political and other issues. Due to this, the implementation of our project not only helps in controlling one of the major modern age social problem but also does it in a fast and effective manner.

2.2.3 Safety

Our project is primarily based on the dataset and the model. The model is executed on Jupyter notebook which is an open-source web application that is used for creating and sharing documents. It is quite secure as it restricts access to the jupyter notebook server and also uses token authentication to keep our data secure.

And when it comes to securing the dataset, we can store it in encrypted folders so that not everyone has access to the data and can make changes in the file. Thus, making it safe.

2.2.4 Ethics

The solution we are implementing doesn't cause harm to any people either physically or virtually. It is quite safe for execution. The user can secure the data using login process for Jupyter notebook application so that not everyone can use the model. This makes our project more reliable and protected.

2.2.5 Cost

There will be very minute requirement for cost maintenance and usage as the data is borrowed from the internet and no extra effort is required to maintain the notebook application as it runs on its own server. All it requires is a PC/laptop with good internet connection.

2.2.6 Type

Our solution comes under the category of standalone product as it just requires a local machine to execute the code and this product has a great potential to solve fake news related problems with ease.

2.2.7 Standards

We have followed the Agile Methodology to create our project. Initially, we gathered all the requirements that are necessary for the project. We have specifically defined the purpose of the software and have divided the entire project into small builds. These builds were worked on for days together to produce an efficient solution for each stage.

Both the design and the testing processes were simultaneous so that it would be easier for use to identify if there was any problem with the code. As soon as there were any changes, we would implement the idea and then repeat the process. This proved out to be an efficient solution that not only made the project execution faster but also easier.

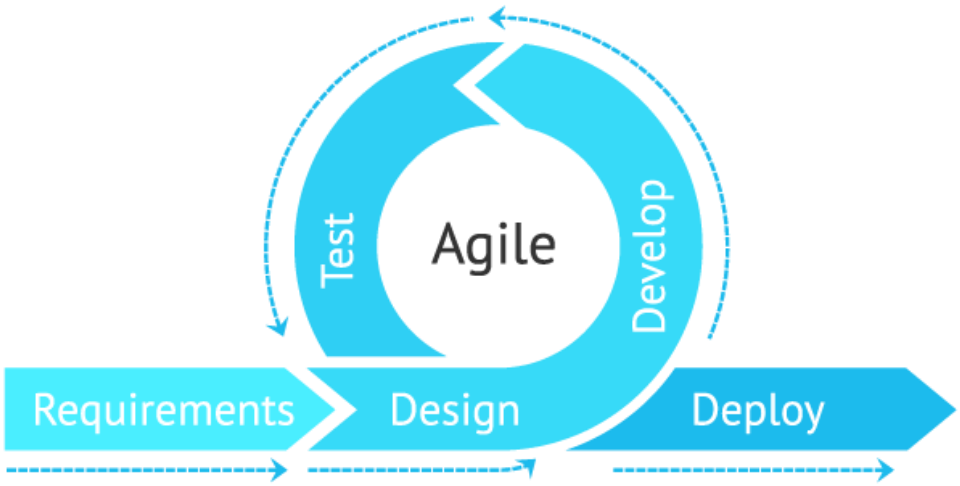

Figure 2.2.7: Agile Model

2.3 SCOPE OF THE PROJECT

The main scope of the project is to create a simple yet effective solution for classifying the news articles. The present systems that are used are not accurate. This has resulted in spread of fake news articles. So, by making use of NLP and machine learning we have decided to bring a solution to this problem that is updated and comparatively much accurate than the already existing solution.

2.4 MODULES DESCRIPTION

There were quite a few machine learning libraries that were used in our project. They were of great use when it came to predicting exactly what we wanted and were very precise too. They are listed down here.

NumPy

NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

Using NumPy, a developer can perform operations like, Mathematical and logical operations on arrays, Fourier transforms and routines for shape manipulation and operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

Pandas

Pandas is a high-level data manipulation tool that is built on the Numpy package and its key data structure is called the Data Frame. Data Frames allow you to store and manipulate tabular data in rows of observations and columns of variables.

Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyse. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Scikit-learn

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

It's also a fantastic library for beginners because it offers a high-level interface for many tasks (e.g. pre-processing data, cross-validation, etc.). This allows you to better practice the entire machine learning workflow and understand the big picture.

Regular expression (or RE)

Regular expression specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

2.5 SYSTEM CONFIGURATION

System analysis is done to know about the requirements in both functional and non-functional perspective. It gives software requirement specification that gives an overview of the system.

Software Requirements

Applications: Jupyter Notebook, Anaconda Python or Google Colab

Language: Python

Tools: Microsoft Excel (For dataset)

Hardware Requirements

RAM: 4GB or more

Processor: 2 GHz or more

3. LITERATURE OVERVIEW

3.1 Machine Learning

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyse the impact of machine learning processes.

3.1.1 Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labelled by humans, and unsupervised learning which provides the algorithm with no labelled data in order to allow it to find structure within its input data. Let's explore these methods in more detail.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labelled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabelled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labelled as fish and images of oceans labelled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabelled shark images as fish and unlabelled ocean images as water.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

Unsupervised Learning

In unsupervised learning, data is unlabelled, so the learning algorithm is left to find commonalities among its input data. As unlabelled data are more abundant than labelled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows

the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

Approaches

As a field, machine learning is closely related to computational statistics, so having background knowledge in statistics is useful for understanding and leveraging machine learning algorithms.

For those who may not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables. Correlation is a measure of association between two variables that are not designated as either dependent or independent. Regression at a basic level is used to examine the relationship between one dependent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities.

3.1.2 Machine Learning Models

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Once you have trained the model, you can use it to reason over data that it hasn't seen before, and make predictions about those data. For example, let's say you want to build an application that can recognize a user's emotions based on their facial expressions. You can train a model by providing it with images of faces that are each tagged with a certain emotion, and then you can use that model in an application that can recognize any user's emotion.

Usually, machine learning models require a lot of data in order for them to perform well. Usually, when training a machine learning model, one needs to collect a large, representative sample of data from a training set. Data from the training set can be as varied as a corpus of text, a collection of images, and data collected from individual users of a service. Overfitting is something to watch out for when training a machine learning model.

A machine learning model is the output of the training process and is defined as the mathematical representation of the real-world process. The machine learning algorithms find the patterns in the training dataset which is used to approximate the target function and is responsible for the mapping of the inputs to the outputs from the available dataset.

These machine learning methods depend upon the type of task and are classified as Classification models, Regression models, Clustering, Dimensionality Reductions, Principal Component Analysis etc.

Regression Technique

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between

variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable based on the value of one or more multiple predictor values. Briefly, the goal is to build a mathematical equation that defines y as a function of the x variables. This equation is then used to predict the outcome (y) on the basis of new values of predictor values (x).

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modelling, and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum". The distance between datapoints and line tells whether a model has captured a strong relationship or not.

3.2 Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. It’s uses include data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning etc.

Jupyter notebooks basically provide an interactive computational environment for developing Python based Data Science applications. They are formerly known as ipython notebooks. The following are some of the features of Jupyter notebooks that makes it one of the best components of Python ML ecosystem –

- Jupyter notebooks can illustrate the analysis process step by step by arranging the stuff like code, images, text, output etc. in a step-by-step manner.
- It helps a data scientist to document the thought process while developing the analysis process.
- One can also capture the result as the part of the notebook.
- With the help of jupyter notebooks, we can share our work with a peer also.

Each .ipynb file is a text file that describes the contents of your notebook in a format called JSON. Each cell and its contents, including image attachments that have been converted into strings of text, is listed therein along with some metadata. You can edit this yourself — if you know what you are doing! — by selecting "Edit > Edit Notebook Metadata" from the menu bar in the notebook. You can also view the contents of your notebook files by selecting "Edit" from the controls on the dashboard

However, the key word there is can. In most cases, there's no reason you should ever need to edit your notebook metadata manually.

There are two fairly prominent terms that you should notice, which are probably new to you: cells and kernels are key both to understanding Jupyter and to what makes it more than just a word processor. Fortunately, these concepts are not difficult to understand.

- A kernel is a "computational engine" that executes the code contained in a notebook document.
- A cell is a container for text to be displayed in the notebook or code to be executed by the notebook's kernel.

3.2.1 Cells

Cells form the body of a notebook. In the screenshot of a new notebook in the section above, that box with the green outline is an empty cell. The first cell in a new notebook is always a code cell.

The output of a code cell also forms part of the document, which is why you can see it in this article. You can always tell the difference between code and Markdown cells because code cells have that label on the left and Markdown cells do not.

The “In” part of the label is simply short for “Input,” while the label number indicates when the cell was executed on the kernel — in this case the cell was executed first. Run the cell again and the label will change to In [2] because now the cell was the second to be run on the kernel. It will become clearer why this is so useful later on when we take a closer look at kernels. From the menu bar, click Insert and select Insert Cell Below to create a new code cell underneath your first and try out the following code to see what happens.

3.2.2 Markdown

Markdown is a lightweight, easy to learn markup language for formatting plain text. Its syntax has a one-to-one correspondence with HTML tags, so some prior knowledge here would be helpful but is definitely not a prerequisite. When attaching images, you have three options:

- Use a URL to an image on the web.
- Use a local URL to an image that you will be keeping alongside your notebook, such as in the same git repo.
- Add an attachment via “Edit > Insert Image”; this will convert the image into a string and store it inside your notebook .ipynb file. Note that this will make your .ipynb file much larger!

There is plenty more to Markdown, especially around hyperlinking, and it’s also possible to simply include plain HTML.Behind every notebook runs a kernel. When you run a code cell, that code is executed within the kernel. Any output is returned back to the cell to be displayed. The kernel’s state persists over time and between cells — it pertains to the document as a whole and not individual cells. For example, if you import libraries or declare variables in one cell, they will be available in another.

Most of the time when you create a notebook, the flow will be top-to-bottom. But it’s common to go back to make changes. When we do need to make changes to an earlier cell, the order of execution we can see on the left of each cell, can help us diagnose problems by seeing what order the cells have run in.

And if we ever wish to reset things, there are several incredibly useful options from the Kernel menu:

- Restart: restarts the kernel, thus clearing all the variables etc that were defined.
- Restart & Clear Output: same as above but will also wipe the output displayed below your code cells.
- Restart & Run All: same as above but will also run all your cells in order from first to last.

If your kernel is ever stuck on a computation and you wish to stop it, you can choose the Interrupt option.

3.2.3 Sharing Your Notebooks

When people talk about sharing their notebooks, there are generally two paradigms they may be considering. Most often, individuals share the end-result of their work, much like this article itself, which means sharing non-interactive, pre-rendered versions of their notebooks. However, it is also possible to collaborate on notebooks with the aid of version control systems such as Git or online platforms like Google Colab.

3.2.4 Exporting Your Notebooks

Jupyter has built-in support for exporting to HTML and PDF as well as several other formats, which you can find from the menu under "File > Download As."

If you wish to share your notebooks with a small private group, this functionality may well be all you need. Indeed, as many researchers in academic institutions are given some public or internal webspace, and because you can export a notebook to an HTML file, Jupyter Notebooks can be an especially convenient way for researchers to share their results with their peers. But if sharing exported files doesn't cut it for you, there are also some immensely popular methods of sharing .ipynb files more directly on the web.

4. ALGORITHM USED FOR FAKE NEWS DETECTION

Our proposed system makes use of Machine Learning algorithms for classifing news articles. The dataset we have included consists of various news articles both real and fake news. The workflow reads the data and trains a regression model based on various attributes. The model is then used to predict the correctness of the article. The proposed system focuses on achieving high accuracy which helps people to follow real news.

4.1 RANDOM FOREST ALGORITHM

Random forest is a supervised learning algorithm which is used for both classification as well as regression. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

4.1.1 Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps −

Step 1 − First, start with the selection of random samples from a given dataset.

Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 − In this step, voting will be performed for every predicted result.

Step 4 − At last, select the most voted prediction result as the final prediction result.

The following diagram will illustrate its working

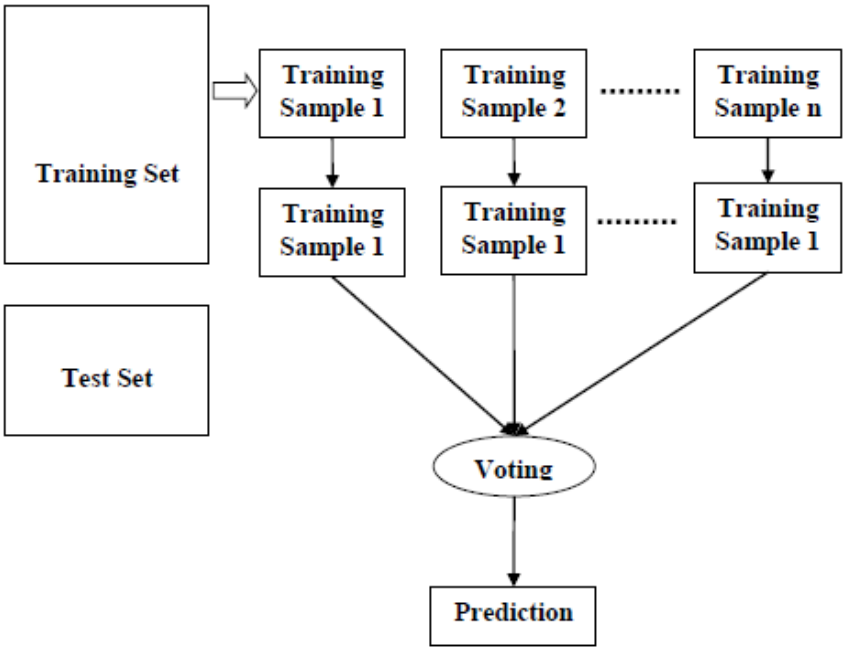

Figure 4.2 Random Forest Algorithm

4.1.2 Benefits of Random Forest Algorithm

1. It can be used for both classification and regression tasks.

2. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model.

3. The classifier of Random Forest can handle missing values.

4. The last advantage is that the Random Forest classifier can be modeled for categorical values.

5. SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

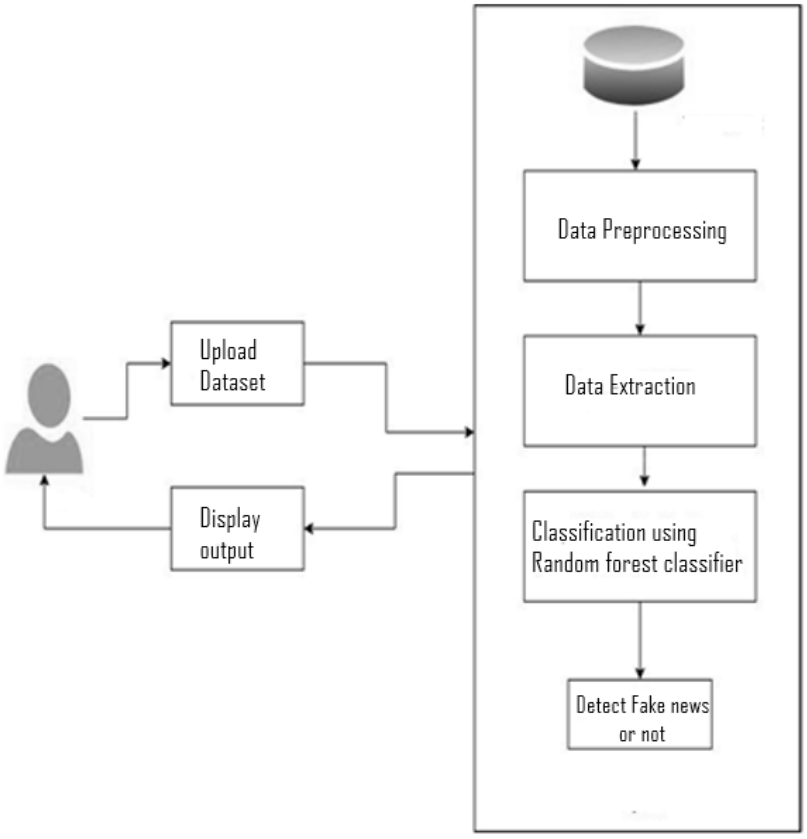

Figure 5.1: System Architecture for a smart system for fake news detection

5.2 UML DIAGRAMS

5.2.1 Component Diagram

Component diagrams are used in modelling the physical aspects of object-oriented systems that are used for visualizing, specifying, and documenting component-based systems.

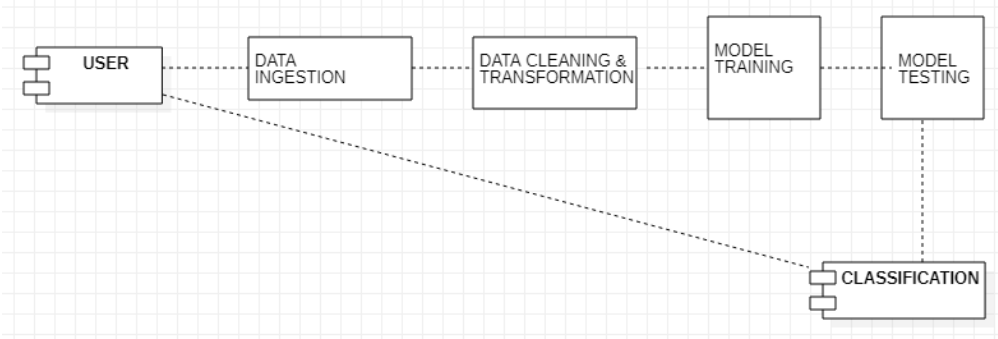

Figure 5.2.1: Component Diagram for a smart system for fake news detection

5.2.2 Deployment Diagram

Deployment diagram shows the execution architecture of a system including nodes such as hardware or software execution environments and the middleware connecting to them.

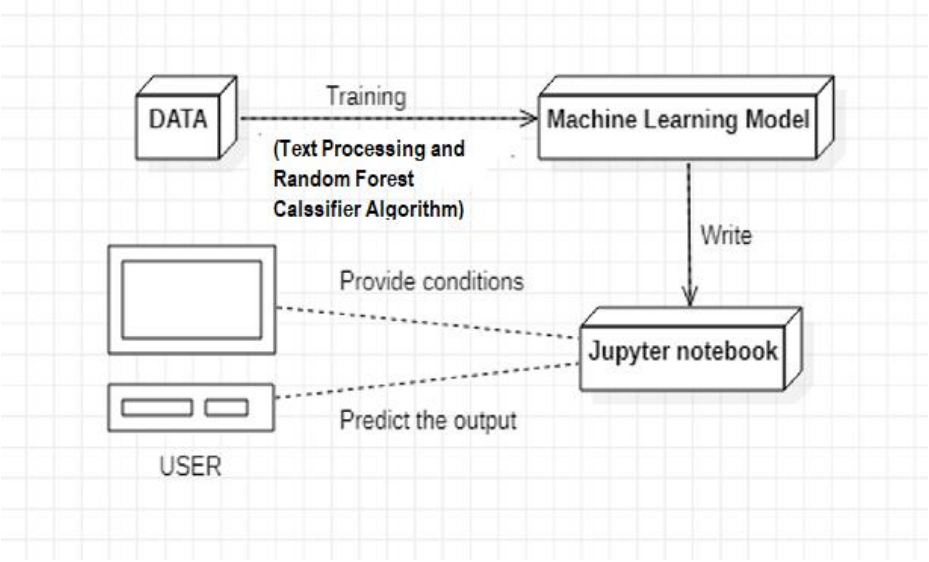

Figure 5.2.2: Deployment Diagram for a smart system for fake news detection

5.2.3 Activity Diagram

Activity diagram describes the flow of control in a system. It consists of activities and links. The flow can be sequential, concurrent, or branched.

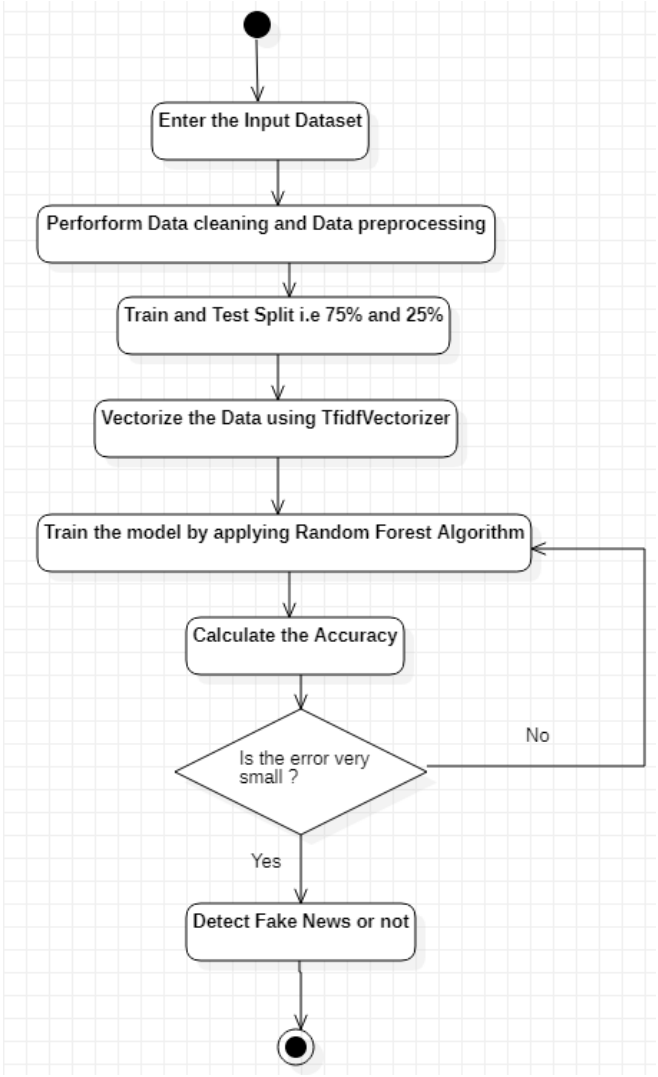

Figure 5.2.3: Activity Diagram for a smart system for fake news detection

5.2.4 Class Diagram

Class diagram consists of classes, interfaces, associations, and collaboration. Class diagrams basically represent the object-oriented view of a system, which is static in nature.

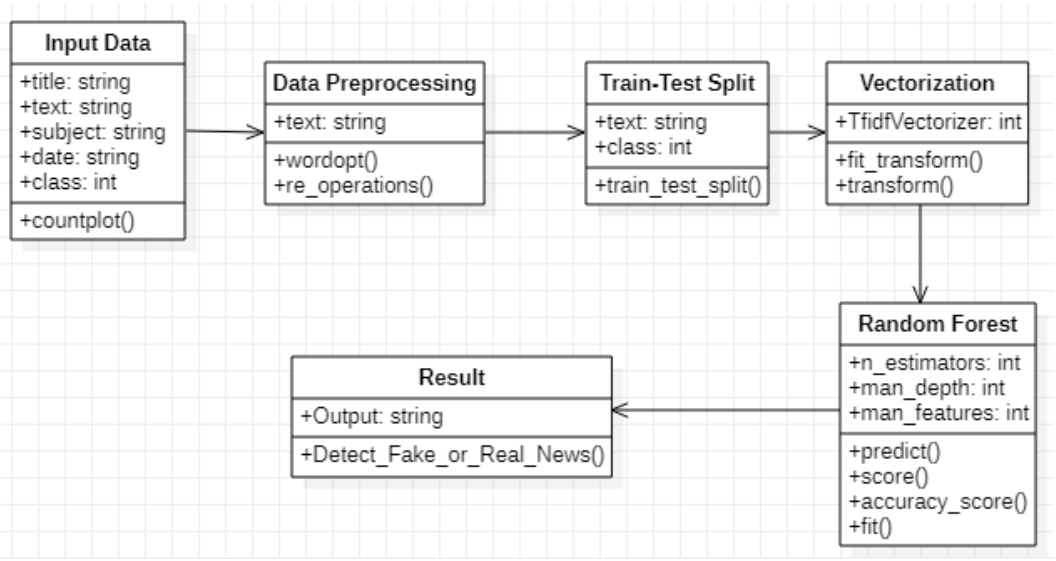

Figure 5.2.4: Class Diagram for a smart system for fake news detection

5.2.5 Use case Diagram

Use case diagrams are a set of use cases, actors, and their relationships. They represent the use case view of a system.

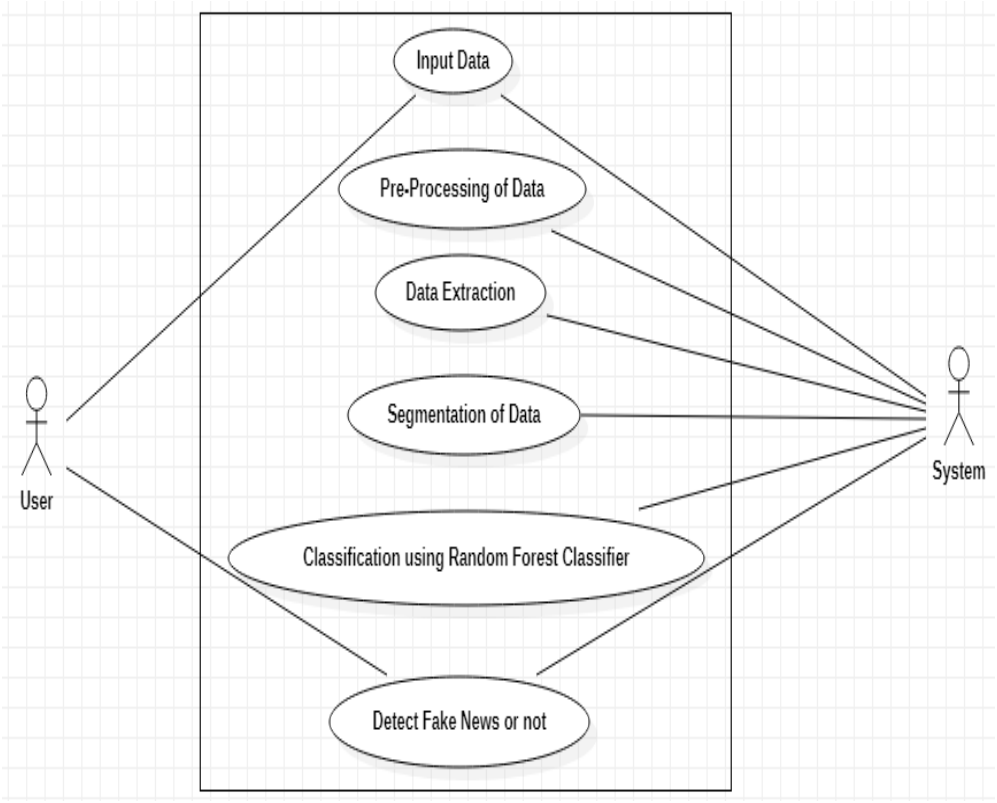

Figure 5.2.5: Use case Diagram for a smart system for fake news detection

5.3 SYSTEM DESIGN

The basic design flow of the system is given below. Initially, the user inputs the weather conditions to the Machine Learning model. This model then performs Data pre-processing and applies the Random Forest Algorithm to the dataset. With the help of this algorithm, we make use of the regression technique to predict the output using the various predictor variables.

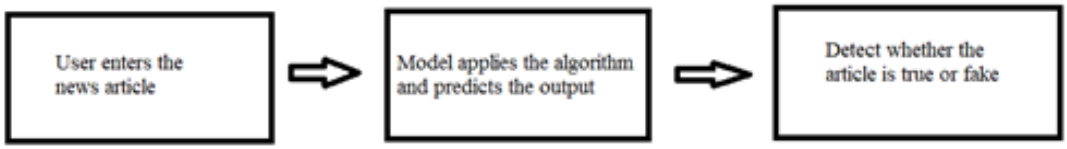

Figure 5.3: System Design for a smart system for fake news detection

6. SAMPLE CODE

6.1 CODING

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report

import re

import string

import nltk

df_false= pd.read_csv("False.csv",encoding='unicode_escape')

df_false['subject'].value_counts()

plt.figure(figsize=(10,6))

sns.countplot(x='subject',data=df_fake)

df_false.head(5)

df_false['text']=df_false['text'].apply(str)

df_true = pd.read_csv("True.csv",encoding='unicode_escape')

df_true['subject'].value_counts()

plt.figure(figsize=(10,6))
```

```
sns.countplot(x='subject',data=df_true)

df_true.tail(5)

df_true['text']=df_true['text'].apply(str)

text=' '.join(df_true['text'].tolist())

df_false["class"] = 0

df_true["class"] = 1

df_false.shape, df_true.shape

df_fake_mt = df_false.tail(10)

for i in range(34750,34770,-1):

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

df_true_mt = df_true.tail(10)

for i in range(23515,23535,-1):

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

df_fake_mt["class"] = 0

df_true_mt["class"] = 1

#manual testing

df_fake_mt.head(5)

df_mt = pd.concat([df_fake_mt,df_true_mt], axis = 0)

df_mt.to_csv("manual_testing.csv")

df_combine = pd.concat([df_fake, df_true], axis =0 )

df_combine.shape
```

```
df = df_combine.drop(["title", "subject","date"], axis = 1)

df=df.dropna()

df = df.sample(frac = 1)

df.reset_index(inplace = True)

df.drop(["index"], axis = 1, inplace = True)

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

df['text'].astype(str)

df['text'] = df['text'].apply(lambda x: " ".join(x.lower() for x in str(x).split() ))

df['text'] = df['text'].apply(wordopt)

x = df["text"]

y = df["class"]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorization = TfidfVectorizer()
xvec_train = vectorization.fit_transform(x_train)
xvec_test = vectorization.transform(x_test)
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor(random_state = 42)
from pprint import pprint
# Look at parameters used by our current forest
print('Parameters currently in use:\n')
pprint(rf.get_params())
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier()
model.fit(xvec_train, y_train)
from pprint import pprint
# Look at parameters used by our current forest
print('Parameters currently in use:\n')
pprint(model.get_params())
RFC = RandomForestClassifier(random_state=0)
RFC.fit(xvec_train, y_train)
pred_rfc = RFC.predict(xvec_test)
RFC.score(xvec_test, y_test)
```

```
print(classification_report(y_test, pred_rfc))
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, pred_rfc)
def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "True News"

def manual_testing(news):
    testing_news = {"text":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xvec_test = vectorization.transform(new_x_test)
    pred_RFC = RFC.predict(new_xvec_test)

    return print("\n\nRFC Prediction: {}".format(output_lable(pred_RFC[0])))

news = str(input("Enter the arcticle: "))
manual_testing(news)
```

7. TESTING

7.1 TESTING

We have done Software Testing to check whether the actual software is matching the expected behaviour and is error free. The execution involved manual tools where we have evaluated the properties of interest and have made sure that there were no errors, gaps or missing requirements in contrast to actual requirements.

This has proved to be helpful as we were able to identify the bugs and errors at an early stage and has helped us ensure reliability, security and high performance which has resulted in saving time and cost effectiveness.

7.2 TYPES OF TESTING

7.2.1 Unit Testing

It focuses on the smallest unit of software design. In this, we test an individual unit or group of interrelated units. It is often done by the programmer by using sample input and observing its corresponding outputs.

7.2.2 Integration Testing

The objective is to take unit tested components and build a program structure that has been dictated by design. Integration testing is testing in which a group of components is combined to produce output.

7.2.3 System Testing

This software is tested such that it works fine for the different operating systems. It is covered under the black box testing technique. In this, we just focus on the required input and output without focusing on internal working.

7.3 TEST CASES

We have implemented various test cases. Few of them are shown below:

```
news = str(input("Enter the arcticle: "))
manual_testing(news)
```

```
Enter the arcticle: Well, here s a Twitter post that backfired on Donald Trump.Over the weekend, billionaire Prince Alwaleed bi
n Talal, who owns a stake in Fox News, along with several other princes and former cabinet officials, were arrested for corrupt
ion by order of Crown Prince Mohammad bin Salman.Ironically, Trump tweeted out his support of the sweeping arrests on Monday.I
have great confidence in King Salman and the Crown Prince of Saudi Arabia, they know exactly what they are doing .  Donald J. T
rump (@realDonaldTrump) November 6, 2017 .Some of those they are harshly treating have been  milking  their country for years!
Donald J. Trump (@realDonaldTrump) November 6, 2017Let that sink in for a minute. Donald Trump, the man who has literally spent
every moment of his adult life milking his own country and continues to do so now from the Oval Office, agrees that people who
do what he does should be treated harshly.Trump and his administration are currently under investigation for colluding with a f
oreign state to undermine our democratic process. Trump is also a tax evader and has made deals with some very shady people inc
luding people with mob and terrorist connections.In fact, Trump s own tax plan would especially benefit him and his family. And
he has placed his own family members in high White House positions, all while using the executive branch to promote his busines
ses, all of which violates ethics rules.Some Twitter users pointed out the irony and hypocrisy of Trump s tweet.Like your team
is doing??? pic.twitter.com/51L89Rrnfs  sweetsallysue (@sweetsallysue) November 7, 2017RELEASE YOUR TAXES and we'll see who has
been "milking" their country for years! #PutinsPuppet #FakePresident  RandeMande (@almondleafer) November 7, 2017Like you have
been milking US for years? #Traitor  Reap what you sow (@2017moderate) November 7, 2017That is a hilarious statement coming fro
m you, that's exactly what you are doing to us.  jbarton (@jlbarton618) November 6, 2017Make no mistake, Trump has milked the U
nited States more than anyone and he intends to milk much more out of taxpayers unless he is removed from office and put in pri
son where he belongs.


RFC Prediction: Fake News
```

Figure 7.3.1 Test Data 1

```
news = str(input("Enter the arcticle: "))
manual_testing(news)
```

```
Enter the arcticle: 21st Century Wire says Our weekly documentary film, curated by our editorial team at 21WIRE.This film docum
ents a series of early scandals and state corruption surrounding Bill Clinton and Hillary Clinton. The film came under heavy cr
iticism in the US media at the time, with Democratic Party officials and mainstream media labelling it as a  partisan hit piec
e.  While Clinton proponents and critics of this film maintain that it s part of a  vast right-wing conspiracy, this film produ
ced by Citizens for Honest Government does cover actual events which took place in and around Arkansas during the Clinton s rei
gn as Governor there, and contains numerous facts and revelations regarding those events  including a massive cocaine smugglin
g operation into Mena, Arkansas under the watch of the Clinton governorship.


RFC Prediction: Fake News
```

Figure 7.3.2 Test Data 2

```
news = str(input("Enter the arcticle: "))
manual_testing(news)
```

```
Enter the arcticle: Egypt s President Abdel Fattah al-Sisi has invited Palestinian President Mahmoud Abbas to Cairo on Monday t
o discuss U.S. President Donald Trump s recognition of Jerusalem as Israel s capital, a presidential statement said on Sunday.
The statement said Sisi wanted to discuss  ways to deal with the crisis in a manner that preserves the rights of the Palestinia
n people and their national sanctities and their legitimate right to establish an independent state with East Jerusalem as its
capital .


RFC Prediction: True News
```

Figure 7.3.3 Test Data 3

8. OUTPUT SCREENS

The Output screens include predicting the area and also finding the error in the prediction and calculating the accuracy.

```
news = str(input("Enter the arcticle: "))
manual_testing(news)
```

```
Enter the arcticle: Coronaviruses are a diverse group of viruses infecting many different animals, and they can cause mild to s
evere respiratory infections in humans. In 2002 and 2012, respectively, two highly pathogenic coronaviruses with zoonotic origi
n, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), emerge
d in humans and caused fatal respiratory illness, making emerging coronaviruses a new public health concern in the twenty-first
century1. At the end of 2019, a novel coronavirus designated as SARS-CoV-2 emerged in the city of Wuhan, China, and caused an o
utbreak of unusual viral pneumonia. Being highly transmissible, this novel coronavirus disease, also known as coronavirus disea
se 2019 (COVID-19), has spread fast all over the world2,3. It has overwhelmingly surpassed SARS and MERS in terms of both the n
umber of infected people and the spatial range of epidemic areas. The ongoing outbreak of COVID-19 has posed an extraordinary t
hreat to global public health4,5. In this Review, we summarize the current understanding of the nature of SARS-CoV-2 and COVID-
19. On the basis of recently published findings, this comprehensive Review covers the basic biology of SARS-CoV-2, including th
e genetic characteristics, the potential zoonotic origin and its receptor binding. Furthermore, we will discuss the clinical an
d epidemiological features, diagnosis of and countermeasures against COVID-19.


RFC Prediction: True News
```

Figure 8.1 Output Screen 1

```
news = str(input("Enter the arcticle: "))
manual_testing(news)
```

```
Enter the arcticle: Coronaviruses are a diverse group of viruses infecting many different animals, and they can cause mild to s
evere respiratory infections in humans. In 2002 and 2012, respectively, two highly pathogenic coronaviruses with zoonotic origi
n, severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), emerge
d in humans and caused fatal respiratory illness, making emerging coronaviruses a new public health concern in the twenty-first
century1. At the end of 2019, a novel coronavirus designated as SARS-CoV-2 emerged in the city of Wuhan, China, and caused an o
utbreak of unusual viral pneumonia. Being highly transmissible, this novel coronavirus disease, also known as coronavirus disea
se 2019 (COVID-19), has spread fast all over the world2,3. It has overwhelmingly surpassed SARS and MERS in terms of both the n
umber of infected people and the spatial range of epidemic areas. The ongoing outbreak of COVID-19 has posed an extraordinary t
hreat to global public health4,5. In this Review, we summarize the current understanding of the nature of SARS-CoV-2 and COVID-
19. On the basis of recently published findings, this comprehensive Review covers the basic biology of SARS-CoV-2, including th
e genetic characteristics, the potential zoonotic origin and its receptor binding. Furthermore, we will discuss the clinical an
d epidemiological features, diagnosis of and countermeasures against COVID-19.


RFC Prediction: True News
```

Figure 8.2 Output Screen 2

9. CONCLUSION

9.1 CONCLUSION

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. The system proposed here is less complex than the already existing system and thus, proves to be very cost effective.

The data we used in our work is collected from the World Wide Web and contains news articles from various domains to cover most of the news rather than specifically classifying a specific category.

9.2 FURTHER ENHANCEMENTS

Through the work done in this project, we have shown that machine learning certainly does have the capacity to pick up on sometimes subtle language patterns that may be difficult for humans to pick up on. The first of aspect that could be improved in this project is augmenting and increasing the size of the dataset. More data would be beneficial in ridding the model of any bias based on specific patterns in the source. Comparing the accuracies with other neural networks and ML algorithms can be used to improve the system.

10. BIBLIOGRAPHY

- https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/
- https://www.yobots.in
- https://www.hindawi.com/journals/complexity/2020/8885861/
- https://en.wikipedia.org/wiki/Detecting_fake_news_online
- https://www.researchgate.net/publication/330849153_A_Hybrid_GAkNNSVM_Algorithm_for_Classification_of_data
- https://www.irjet.net/archives/V6/i4/IRJET-V6I4342.pdf
- https://www.irjet.net/archives/V7/i6/IRJET-V7I6688.pdf