Fake News Detector

*A project submitted in partial fulfillment of the requirements for the award of the degree of*

Bachelor of Technology in

INFORMATION TECHNOLOGY



Submitted by:

Mr. Sourabh Hembrom Roll No. 12012041

Group ID: 119

Supervised By: Dr. Mukesh Mann

# INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, SONEPAT -131201, HARYANA, INDIA

## SELF DECLARATION

I hereby declare that work contained in the project titled “Fake News Detector” is original. I have followed the standards of project ethics to the best of my abilities. I have acknowledged all sources of information which I have used in the project.

Name: Sourabh Hembrom

Roll No.: 12012041

Department of Information Technology,

Indian Institute of Information Technology, Sonepat-131201, Haryana, India.

## CERTIFICATE

This is to certify that **Mr. Sourabh Hembrom** has worked on the project entitled “Fake News Detector” under my supervision and guidance.

The contents of the project, being submitted to the Department of Information Technology, IIIT, Sonepat, for the award of the degree of B.Tech in Information Technology, are original and have been carried out by the candidate himself. This project has not been submitted in full or part for the award of any other degree or diploma to this or any other university.

Dr. Mukesh Mann Supervisor

Department of Information Technology,

Indian Institute of Information Technology, Sonepat-131201, Haryana, India

## Acknowledgement

I would like to express our special thanks of gratitude to our teacher (Dr. Mukesh Mann) who gave me the golden opportunity to do this wonderful project on the topic (Fake News Detector), which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them.

Secondly, I would also like to thank our parents and friends who helped me a lot in finalizing this project within the limited time frame*.*

I am overwhelmed in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.

Any attempt at any level can 't be satisfactorily completed without the support and guidance of my parents and friends.

I would like to thank my parents who helped me a lot in gathering different information, collecting data and guiding me from time to time in making this project, despite of their busy schedules, they gave me different ideas in making this project unique.

Thanking you,

Sourabh Hembrom

12012041

IT

## Abstract

Name of the student: **Sourabh Hembrom**

Roll no: **12012041**

Degree for which submitted: **BTech (IT)**

Department of **Information Technology, IIIT Sonepat.**

Project Title: **Fake News Detector**

Name of the supervisor: **Dr. Mukesh Mann**

Month and year of the submission: **December, 2021**

With our world producing an ever-growing huge amount of data exponentially per second by machines,

there is a concern that this data can be false (or fake). Fake news (or data) can pose many dangers to our

world. Imagine what happens if due to some false information you are given the wrong medicine.

Luckily, this problem can be addressed using machine learning. We can develop a machine learning

model in python which can detect whether the news is fake or not. In this project, we build a classifier

that detects whether the news is fake or not. We pre-process the text data from our dataset using TF-IDF

Vectorizer. We apply passive aggressive classifier algorithm to the pre-processed text then train and

evaluate our model on the dataset.

## LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| NLP | Natural Language Processing |
| ML | Machine Learning |
| UI | User Interface |
| TF | Term Frequency |
| IDF | Inverse Document Frequency |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure’s No.** | **Name of Figures** | **Page**  **Nos.** |
| 1 | *Workflow Diagram* | 19 |
| 2 | *Term frequency formula* | 20 |
| 3 | *Inverse document formula* | 20 |
| 4 | *USE CASE DIAGRAM* | 22 |
| 5 | *Tools for this project* | 23 |
| 6 | *Home page* | 24 |
| 7 | *Check News page* | 25 |
| 8 | *Sign Up page* | 25 |
| 9 | *Flowchart to proceed* | 26 |
| 10 | *Flowchart to sign up* | 27 |
| 11 | *Real news entered in content* | 27 |
| 12 | *Real news* | 28 |
| 13 | *Fake news entered in content* | 29 |
| 14 | *Fake news* | 29 |
| 15 | *Details entered in sign up page* | 30 |
| 16 | *After submitting details in signup page* | 30 |

|  |  |  |
| --- | --- | --- |
| **TABLE OF CONTENTS** | | |
| 1.1 | Introduction | 9 |
| 1.2 | Problem Outline | 10 |
| 1.3 | Project Objectives | 11 |
| 1.4 | Project Methodology | 12 |
| 2.1 | Introduction (Study and review of literature) | 13 |
| 2.2 | Important of detecting fake news | 14 |
| 2.3 | Possible ways to build the project | 15-17 |
| 3.1 | Introduction (Implementation of Fake News Detector) | 18 |
| 3.2 | Method of Implementation | 19 |
| 3.3 | Techniques used for this project | 20-22 |
| 3.4 | Use Case Diagram | 22 |
| 3.5 | Tools used for this project | 23 |
| 3.6 | Frontend development for this project | 24-26 |
| 3.7 | Backend development for this project | 26-27 |
| 3.8 | Test and Results | 27-30 |
| 4.1 | Conclusion | 31 |
| 4.2 | Limitation | 32 |

**1)Introduction**

**1.1 Introduction**

Our project that goes by name “Fake news detector” aims to detect whether the news

is fake or real by implementing NLP (Natural Language Process) concept which is a

part of machine learning.

As a human being when we read a sentence or a paragraph, we can interpret the

words with whole documents and understand the context. Given today’s volume of

news, it is possible to teach a computer how to read and understand the difference

between the real news and the fake news using NLP. The building blocks are datasets

and machine learning algorithms.

**1.2 Problem outline**

Fake news, the name itself explains what it means, fake news is **defined as those**

**news stories that are false, the** story itself isfabricated**, with no verifiable facts,**

**sources, or quotes.** Sometimes these stories may be propaganda that is intentionally

designed to mislead the reader or may be designed as clickbait written for economic

incentives (the writer profits on the number of people who click on the story). In recent

years, fake news stories have proliferated via social media, in part because they are

so easily and quickly shared online.

This is often done to impose certain ideas and is often done with political agendas.

Often fake news will mimic real headline and twist the story. Fake news can influence

people’s perceptions, the rise of fake news has become a global problem that even

major companies like Facebook and Google are struggling to solve. It can be very

difficult to determine whether factual or not without addition context and human

judgement.

**1.3 Project Objective**

The main objective of the project is to find a way to utilize NLP to identify and classify

fake news articles, to detect the fake news, which is a classic text classification

problem. First The dataset containing both fake and real news is collected then the text

in the dataset it is pre-processed after pre-processing the text the whole article is

converted into features for use in supervised model. A model is built through training

with help of train and test data so that the trained model can differentiate between real

news and fake news. So ultimately our end goal is to develop a model that classify the

article either fake or real.

**1.4 Methodology**

In this project we are using machine learning concepts, the steps that we have

followed are as follows:

1)Data collection and preparation: which includes getting the training data, cleaning it

and prepare it to the process of feature extraction.

2)Feature selection: which consists of identifying the features that are most useful for

the problem under examination.

3)Algorithm choice: from the given dataset, after selecting the features, a suitable

algorithm is selected to extract these features from the dataset.

4)Parameter and model selection: which means choosing the machine learning model

and setting its parameter to guarantee the best performance with extracted features.

5)Train the model: which includes the training of the model with the help of training

data from the dataset so it can be ready to predict the output of new data.

6)Testing the model: after training the model, the model is tested with the help of test

data extracted from the collected dataset and accuracy of the prediction is evaluated.

**2) Study and review of literature**

In this section we will talk about the research work related to the project that we found out

while doing our project. In this way, we will try to create project which accomplishes its task

but will also improve its performance gradually by implementing this research methodology.

**2.1 Introduction**

Fake news refers to information content that is false, misleading or whose source cannot be

verified. This content may be generated to intentionally damage reputations, deceive, or to gain

attention. The term rose to popularity during the 2016 US Presidential Elections. It was

reported that fake news likely influenced the results of the elections.

Various types of fake news include:

* Clickbait: Often eye-catching content to capture readers at the expense of being factual.
* Satire/parody: This type of content is considered to be fun and humorous thus considered to be entertaining, yet some readers may interpret the content as fact.
* Propaganda: This is content meant to mislead and influence the reader.
* Biased/partisan/hyper-partisan: Oftentimes this is biased political content claiming to be impartial.
* Unreliable news: Journalists may publish news whose sources are unverified, or without carrying out any form of fact checking themselves.

Social media platforms are incredibly influential. According to internet live stats the estimated

daily number of tweets is about 500 million. These platforms are ubiquitous. They are the go-to

environment to share thoughts, feelings, opinions, and intentions. This provides ideal

conditions to distribute news with minimal guidelines and restrictions.

In today’s world, it is normal to receive news from online sources like social media. News is

often subjective to readers. We often choose to ingest content that appeals to the different

emotions we have. So, considering this, the information that gets the most reach may not be

real or accurate news. Additionally, real news may be twisted in transmission. A reader may

end up with different versions of the same news. This may lead to information overload.

**2.2 Importance of detecting fake news**

At a time when the globe is defined by a pandemic, public health depends on reliable

information. Yet we stare down the barrel of an infodemic. An infodemic is the combination of

the word information and epidemic. It is an excessive amount of information about a problem

that makes the solution more difficult. It also defines a wide and rapid spread of

misinformation.

This means that our individual health is a collective responsibility. It is tied to the behaviour of

other people since news influences the behaviour of the audience. The World Health

Organization has highlighted the dangers of a COVID-19 driven infodemic. It presents as much

danger as the virus itself. According to WHO, fake news spreads faster and more easily than the

virus.

Examples of challenges of such an infodemic include:

* Promoting and selling of fake coronavirus cures.
* Spreading myths and rumours about the nature and spread of the virus.
* Conspiracy theories about the origin and intention of the virus.
* Encouraging unfounded remedies. Some touted remedies are harmless, others are comical, while others can be quite hazardous.

**2.3 Possible ways to build the project**

These two approaches focus on the methods used, as opposed to the content being analysed.

They may also both involve NLP in their methodology.

NLP enables computers to understand natural/human language and respond appropriately.

Hence, there are two aspects involved:

* Natural Language Understanding
* Natural Language Generation

The two approaches to fake news detection are:

* Machine Learning approach
* Deep Learning approach

*Machine Learning Approach*

Machine learning refers to giving computers the ability to learn without explicitly being

programmed. A machine learning approach uses machine learning algorithms to detect

misinformation. Examples of these algorithms include:

**Naïve Bayes**: uses probabilistic approaches based on Bayes theorem. This algorithm is often used for text classification.

**Decision Tree**: a supervised learning algorithm that has a tree-like flow. It helps in decision making. A useful algorithm for both classification and regression tasks.

**Random forest**: simply a combination of decision trees.

**Support Vector Machine**: a supervised learning algorithm. It examines data for classification and regression analysis. It classifies data into two categories.

**Logistic Regression**: contrary to the name, it is a classification algorithm used to estimate discrete values.

**K-nearest-neighbour**: a simple algorithm that is used for both classification and regression tasks. Though it is more widely used for classification problems.

Datasets are used to refine the algorithms. These datasets may be split as training data or test

data. I have come across a lot of research where a system combines various machine learning

algorithms and data mining. This is often carried out on social media platforms, especially

Twitter data. For example, a model may combine machine learning, through Naïve Bayes, Support Vector Machine(SVM), and Natural Language Processing (NLP) to detect fake news.

Naïve Bayes and Support Vector Machine are classification models in this process.

Depending on the nature of the data, the two classifiers can be applied to a dataset and their

performance compared. On the other hand, these classifiers can be used in an ensemble method

to enhance each other’s results in classification tasks, therefore improving model accuracy. As

mentioned above, Naïve Bayes is popular in text classification tasks therefore it’s considered

for such tasks often.

SVM classifies data into two categories. In the context of fake news detection, these categories

are likely to be “true” or “false”. It is also an algorithm that works well on semi-structured

datasets and is very adaptable. Pairing SVM and Naïve Bayes is therefore effective for fake

news detection tasks.

NLP may play a role in extracting features from data. It may also come in handy when

attempting to contextualize text data since this is not a strong suit of traditional machine

learning models. NLP may also be utilized through sentiment analysis of the data, given

sentiment analysis is a subfield of NLP.

Sentiment analysis is the process of deriving meaningful patterns in text data. It can provide

information on the context of data by describing the sentiments of a given population.

The accuracy of the results is usually determined by the combinations of models used and the

datasets involved. A combination of available toolkits with Bayesian learning may be used to

develop a fake news detector. These toolkits include Textblob, Natural Language, and SciPy.

But a challenge exists with some of these traditional machine learning approaches. They treat

fake news detection as a binary classification task. These models alone struggle to contextualize

text data. They need structured/labelled data. In fact, machine learning models struggle to solve

complex queries with huge amounts of data. This is where deep learning models come into

play.

*Deep Learning Approach*

Deep learning algorithm’s function is similar to machine learning algorithms. But there is a key

difference. Deep learning algorithms have layers that interpret data differently. Artificial neural

networks refer to the network of such algorithms.

Purely deep learning perspectives towards fake news detection have been explored in many

cases.

A methodology may involve building classifiers to predict the validity of news based only on

news content. This may be achieved using Recurrent Neural Network (RNN) models and long-

short term memories (LSTM).

RNN is a neural network containing loops that allow information to be stored within the

network. Previous experiences influence upcoming events in RNNs. The storage of information

can be attributed to LSTM. LSTM refers to artificial recurrent neural networks that allow

information to persist within them. They are the building blocks for RNN layers. LSTM units

provide the ability to “recall” values over a time interval. This influences the relationships

between words and their occurrences.

A combination of machine learning and deep learning techniques is feasible. There are many

published works that combine the two. The aim is not only to detect fake news, but to also

achieve the highest possible accuracy levels in the detection.

**3)Implementation of Fake News Detector**

**3.1 Introduction**

Ever read a piece of news which just seems bogus? We all encounter such news articles, and

instinctively recognise that something doesn’t feel right. Because of so many posts out there, it

is nearly impossible to separate the right from the wrong. Here, we are not only talking about

spurious claims and the factual points, but rather, the things which look wrong intricately in the

language itself.

There are two ways of claiming that some news is fake or not: First, an attack on the factual

points. Second, the language. The former can only be done through substantial searches into the

internet with automated query systems. It could be an overwhelming task, especially for

someone who is just getting started with data science and natural language processing.

The latter is possible through a natural language processing followed by a machine learning

implementation. It is how we implement our fake news detector project. The current chapter

highlights the techniques , methods and tools that are used for the development of the project.

From development of the front end of the application to the backend process of the application

is covered in this chapter including the screenshots of the application in action.

**3.2 Method of Implementation**

Diagram

Description automatically generated

*Fig (1): Workflow Diagram*

This above figure is the workflow diagram of the project in which:

i)This first is to collect the dataset. The collected dataset should be accurate, so we have

collected dataset from reliable site called Kaggle.

ii)Then comes the pre-processing of the data which involves removing of stop words,

stemming, tokenization etc.

iii)Then we move to feature extraction which is a method in which pre-processed data is

converted into features to reduce the amount of useless data from dataset.

iv)After feature extraction, we split the data into two parts that is train data and test data.

v) We use the train data to train the selected model or algorithm which can give best

performance with the extracted features to get the trained model.

vi) Now we use the test data to check how the trained model is working by checking its

accuracy obtained during the test.

vii)After testing the model our final model is ready.

**3.3 Techniques used for this project**

*TF-IDF vectorizer for feature extraction*

It is used to convert data into matrix of features making it is easier for machine to understand

the main context since machine learning algorithm cannot work on raw data. It highlights a

specific value which might not be too frequent in data but holds great importance. TF-IDF

value of the word increase as the total number of times that word appear in the sentence, but it

also decreases with the number of sentences in which the word appears. It is composed of two

parts Term Frequency(TF) and Inverse Document Frequency(IDF).

*Term Frequency formula*

Graphical user interface, text

Description automatically generated

*Fig (2): Term frequency formula*

*Inverse Document Formula*

Table

Description automatically generated

*Fig (3): Inverse document formula*

Now for calculating TF-IDF formula is as follow:

*TF-IDF = TF × IDF*

High TF-IDF value of a word means the word is rare but has high importance.

*Passive Aggressive algorithm for training the model*

It is one of the few ‘online-learning algorithms‘. In online machine learning algorithms, the

input data comes in sequential order and the machine learning model is updated step-by-step, as

opposed to batch learning, where the entire training dataset is used at once. This is very useful

in situations where there is a huge amount of data, and it is computationally infeasible to train the entire dataset because of the sheer size of the data. We can simply say that an online-

learning algorithm will get a training example, update the classifier, and then throw away the

example.

*Pickle module for storing python objects*

Python pickle module is used for serializing and de-serializing python object structures. The

process to converts any kind of python objects (list, dictionary, etc.) into byte streams (0s and

1s) is called pickling or serialization or flattening or marshalling. We can converts the byte

stream (generated through pickling) back into python objects by a process called as

unpickling.

In real world scenario, the use pickling and unpickling are widespread as they allow us to

easily transfer data from one server/system to another and then store it in a file or database.

Pickling is mostly useful in Machine Learning. A machine learning model is trained on a very

large dataset, and training a model consumes a substantial amount of time. So if we have to

train the same model, it would not be a good call again and again. To avoid or lessen the time

and hard work, pickling is highly useful. We have to train our model just once, which then can

be saved into a local disk, and when we need to test our model, we can just load it from the disk

without having to train it again.

*Scikit-Learn for importing necessary modules for implementation of the project*

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It

provides a selection of efficient tools for machine learning and statistical modelling including

classification, regression, clustering and dimensionality reduction via a consistence interface in

Python. It provides modules for loading the dataset, splitting of the dataset, pre-processing of

the data, extraction of features, importing required machine learning algorithm and modules to

check the accuracy score of the model.

**3.4 Use Case Diagram**

Diagram

Description automatically generated

#### Fig(4): USE CASE DIAGRAM

**THE FIGURE SHOWS USE CASE DIAGRAM OF WEBSITE.**

**3.5 Tools used for this project**

Graphical user interface

Description automatically generated

#### Fig(5): Tools for this project

**3.6 Frontend development for this project**

The frontend development was done using python flask, HTML , CSS, and JavaScript. There

are three pages in the user interface Home, Check News and Sign Up.

*i)Home page*

Graphical user interface, text

Description automatically generated

#### Fig(6): Home page

The frontend page of the application is simple and easy to use. The Home page containing a

background image shows up with a text asking is the news real and there is a button named

check here, clicking it will help you reach the main page where you can make use of the

project. On top there are navigation bars as well through which you can directly jump into

different page.

*ii)Check News page*

Graphical user interface, text, application, chat or text message

Description automatically generated

#### Fig(7): Check News page

It is the Check News page where the main task is handled. It contains Text area field where you

can the content of the news you want to check. It contains two buttons named predict and reset.

When news content is entered in the text area the predict is then clicked to check whether the

news is real or fake. The reset button on the other hand is used to refresh the page.

*iii)Sign up page*

Graphical user interface, application

Description automatically generated

#### Fig(8): Sign Up page

This contains 4 input field that is name, date of birth, mobile number, and email. These data are

collected from the user and is saved in a database future reference and their feedback. When

submit button is clicked the data is collected and a thank you message is displayed.

**3.7 Backend development for this project**

For backend we have used Jupyter notebook and python. When the user enter the news content

the check news page the model does the prediction whether the news is fake or real.

Diagram

Description automatically generated

#### Fig(9): Flowchart to proceed

Then in sign up page user details can be saved in a database and can used for future reference or

reviews.

Diagram

Description automatically generated

#### Fig(10): Flowchart to sign up

**3.8 Test and Results**

*When the news is real*

i)Entered the news to content

Graphical user interface, text, application, chat or text message

Description automatically generated

#### Fig(11): Real news entered in content

ii)Clicked on predict button.

Graphical user interface, text, application, chat or text message

Description automatically generated

#### Fig(12): Real news

This is the result.

*When the news is fake*

i)Entered the news to content

Graphical user interface, text, application, chat or text message

Description automatically generated

#### Fig(13): Fake news entered in content

ii)Clicked on predict button.

Graphical user interface

Description automatically generated

#### Fig(14): Fake news

This is the result.

*Sign up Page*

i)Entered user details

Graphical user interface, application

Description automatically generated

#### Fig(15): Details entered in sign up page

ii)Clicked on Submit button

A picture containing text

Description automatically generated

#### Fig(16): After submitting details in signup page

Then a Thank you message pops up on the screen. User can then return to any page of his wish

by clicking on the desired tab on navigation bar.

**4)Conclusion**

**4.1 Conclusion**

The way fake news is adapting technology, better and better processing models would be

required. And these models would be more into natural language understanding and less posed

as a machine learning model itself. The models can also be fine-tuned according to the features

used. The dataset could be made dynamically adaptable to make it work on current data. But

that would require a model exhaustively trained on the current news articles.

So, if more data is available, better models could be made and the applicability of fake news

detection projects can be improved. But the internal scheme and core pipelines would remain

the same. The pipelines explained are highly adaptable to any experiments you may want to

conduct.

In this python is used for building fake news detection projects because of its dynamic typing,

built-in data structures, powerful libraries, frameworks, and community support. The other

requisite skills required to develop a fake news detection project in Python are Machine

Learning, Natural Language Processing, and Artificial Intelligence. But here we have used a

very basic machine learning method to build this project using python.

Since most of the fake news is found on social media platforms, segregating the real and fake

news can be difficult. Social media platforms and most media firms utilize the Fake News

Detection Project to automatically determine whether or not the news being circulated is

fabricated. Most companies use machine learning in addition to the project to automate this

process of finding fake news rather than relying on humans to go through the tedious task.

**4.2 Limitation**

The project aims to detect the news and predict whether it is fake or real, but it is still not

perfect. In this unit we have mentioned few limitations of this project. First thing is that it needs

a very huge amount of accurate dataset only then it shows results close to perfect, in this project

the dataset that contains the news is very limited so any news that is not at all related to the

news from the collected in dataset, it will predict the result as fake even if the news is real, in

other words in this project the correct prediction can be only done only if the news is somewhat

related to that of dataset. This can be corrected only after collecting lots of correct data which

will take time to develop. Second limitation is the accuracy scored by this project is 93 percent,

which means it is not 100 percent accurate one the reason is that algorithm is not perfect

enough, different algorithm for different datasets gives different accuracy score, in this project

algorithm which we have used is not yet perfect. To solve this, we can combine different

methods and algorithm to increase the accuracy but will take many experiments in order to find

the right methods.