



**Module Code & Module Title**

**CU6051NP Artificial Intelligence**

**25% Individual Coursework**

**Submission: Proposal**

**Academic Semester: Autumn Semester 2025**

**Credit: 15 credit semester long module**

**Student Name: Kripa Pandey**

**London Met ID: 23048816**

**College ID: NP04CP4A230022**

**Assignment Due Date: 17/12/2025**

**Assignment Submission Date: 17/12/2025**

**Submitted To: Mr. Jeevan Prakash Pant**

|             |  |
|-------------|--|
| GitHub Link |  |
|-------------|--|

*I confirm that I understand my coursework needs to be submitted online via MST Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.*

## Table of Contents

|   |           |
|---|-----------|
| <b>1. Introduction</b>  | <b>4</b>  |
| <b>1.1. Similar Datasets</b>  | <b>5</b>  |
| <b>1. LIAR Dataset (2017)</b>   | <b>5</b>  |
| <b>2. Fake Newsnet Dataset</b>  | <b>6</b>  |
| <b>3. Kaggle Fake News Dataset</b>                                      | <b>7</b>  |
| <b>4. Facebook's Automated Fact-Checking</b>                            | <b>8</b>  |
| <b>5. Google's Fact-Check Tools</b>                                     | <b>9</b>  |
| <b>6. Twitter/X Community Notes</b>                                     | <b>9</b>  |
| <b>7. NewsGuard</b>   | <b>10</b> |
| <b>2. Background</b>  | <b>11</b> |
| <b>2.1. Research Work Done on the Fake News Problem</b>                 | <b>11</b> |
| <b>2.2. Review and Analysis of the Related Work</b>                     | <b>14</b> |
| <b>3. Solutions</b>   | <b>18</b> |
| <b>3.1. Explanation of the proposed solution to solving the problem</b> | <b>18</b> |
| <b>3.2. Explanation of the AI algorithms used</b>                       | <b>20</b> |
| <b>A. Natural Language Processing (NLP)</b>                             | <b>20</b> |
| <b>B. Machine Learning (ML)</b>   | <b>22</b> |
| <b>3.3. Pseudocode of the Solution</b>                                  | <b>27</b> |
| <b>3.3.1. Pseudocode of Fake News Detection System</b>                  | <b>28</b> |
| <b>3.4. Diagrammatical representations of the solution</b>              | <b>30</b> |
| <b>4. Background</b>  | <b>31</b> |
| <b>4.1. Analysis of the work data</b>                                   | <b>31</b> |
| <b>4.2. How the Solution Works for Real-World Situations</b>            | <b>32</b> |
| <b>4.3. Further Works</b>   | <b>33</b> |
| <b>References</b>   | <b>35</b> |

## Table of Figures

|  |    |
|--|----|
| Figure 1: Liar Dataset .....                       | 5  |
| Figure 2: Fake Newsnet Dataset.....                | 6  |
| Figure 3: Kaggle.....                              | 7  |
| Figure 4: Facebook's Automated Fast-Checking ..... | 8  |
| Figure 5: Google's Fact-Check Tools.....           | 9  |
| Figure 6: Twitter/X Community Notes .....          | 9  |
| Figure 7: NewsGuard.....                           | 10 |
| Figure 8: Flowchart .....                          | 30 |

## 1. Introduction

In today's digital world, information travels very fast through social media, online news sites, and messaging apps. This fast sharing has many advantages but brings one serious problem: the spread of fake news. Fake news means false or misleading information made to look like real news. It can change people's opinions, affect elections, create fear, and reduce trust in real new sources. Since it is impossible for humans to manually check every piece of news, we need an automatic system that can quickly and precisely detect fake news.

This project focuses on the use of Machine Learning to help solve this problem. Machine Learning, especially supervised learning, is usually used for classification tasks, where we need to separate data into different groups. We have two kinds of groups in fake news detection: Fake and Real. We train an ML model with a dataset containing news articles labeled as real or fake; hence, it learns the patterns and will make decision whether a new article is trustworthy or not.

The key tasks of this coursework are investigating the current approaches in the field of fake news detection, understanding how machine learning helps in the respect, and designing a simple machine learning based solution, this includes steps like data cleaning, extraction of useful features, model selection, testing for its accuracy, and finally using it for predicting new articles as real or fake. To make the explanation clear diagrams and pseudocode are also included (Mao, 2025).

## 1.1. Similar Datasets

Many fake news detection systems and datasets have already been developed. Below are some of the most widely used research works and projects in this field.

### 1. LIAR Dataset (2017)

**Table 1.** LIAR Dataset samples distribution.

|                    | Training | Validation | Testing |
|--------------------|----------|------------|---------|
| <b>Pants-fire</b>  | 839      | 116        | 92      |
| <b>FALSE</b>       | 1994     | 263        | 249     |
| <b>Barely-true</b> | 1654     | 236        | 212     |
| <b>Half-true</b>   | 2114     | 248        | 265     |
| <b>Mostly-true</b> | 1962     | 251        | 241     |
| <b>TRUE</b>        | 1676     | 169        | 207     |
| <b>Total</b>       | 10,239   | 1283       | 1266    |

*Figure 1: Liar Dataset*

The LIAR dataset is one of the most popular benchmark datasets for fake news detection. It contains around 12,800 labelled statements from political debates, speeches, and social media (Wang, 2017). Each statement is labelled into categories such as:

- True
- Mostly-true
- Half-true
- Barely-true
- False
- Pants-fire

#### **Contribution:**

It supports multi-class classification and helps train ML models to detect subtle levels of truthfulness.

#### **Limitation:**

It focuses mostly on political statements.

## 2. Fake Newsnet Dataset

| Platform         | BuzzFeed | PolitiFact |
|------------------|----------|------------|
| # Users          | 15,257   | 23,865     |
| # Engagements    | 25,240   | 37,259     |
| # Social Links   | 634,750  | 574,744    |
| # Candidate news | 182      | 240        |
| # True news      | 91       | 120        |
| # Fake news      | 91       | 120        |
| # Publisher      | 9        | 91         |

*Figure 2: Fake Newsnet Dataset*

Fake Newsnet is a large-scale dataset which includes:

- News content
- Social network postings
- User profiles
- Engagement patterns including likes, shares, retweets.

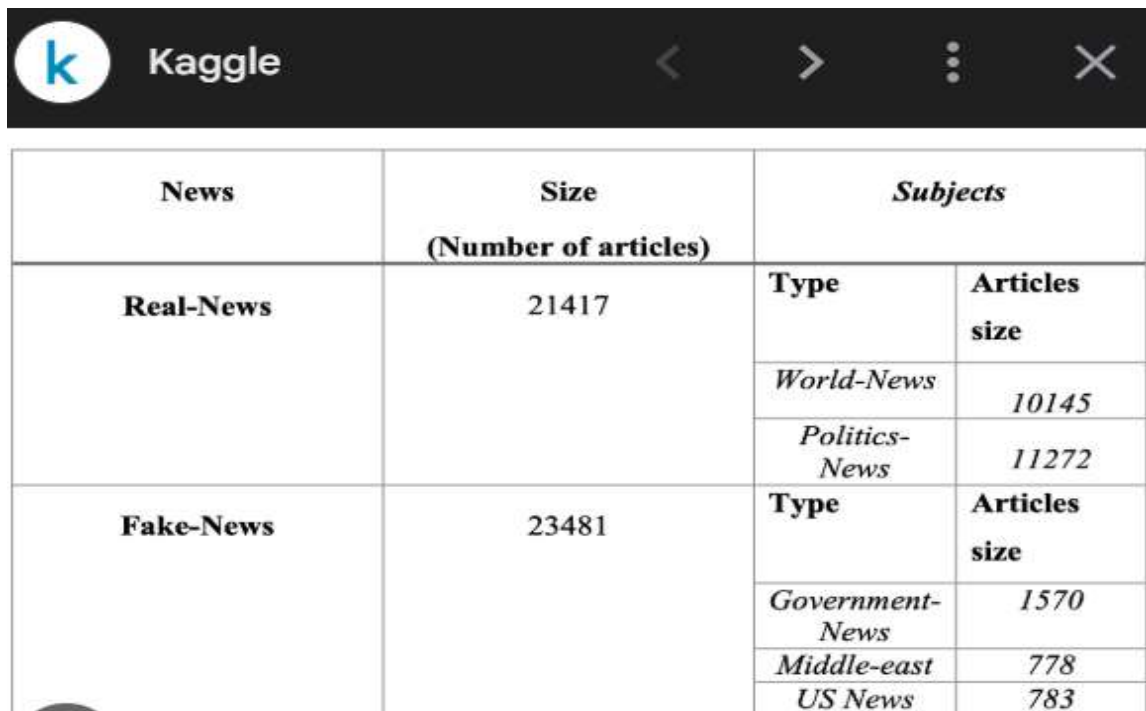
### **Contribution:**

It supports advanced detection approaches that combine text analysis with user behaviour analysis (Sastrawan, 2021).

### **Limitation:**

The dataset is complex, and expert technical skills are called for.

### 3. Kaggle Fake News Dataset



| News      | Size<br>(Number of articles) | Subjects        |               |
|-----------|------------------------------|-----------------|---------------|
| Real-News | 21417                        | Type            | Articles size |
|           |                              | World-News      | 10145         |
|           |                              | Politics-News   | 11272         |
| Fake-News | 23481                        | Type            | Articles size |
|           |                              | Government-News | 1570          |
|           |                              | Middle-east     | 778           |
|           |                              | US News         | 783           |

Figure 3: Kaggle

This dataset contains thousands of real and fake articles with labels. Many students and researchers use it for:

- Binary classification - Real vs Fake
- NLP preprocessing
- ML model comparison

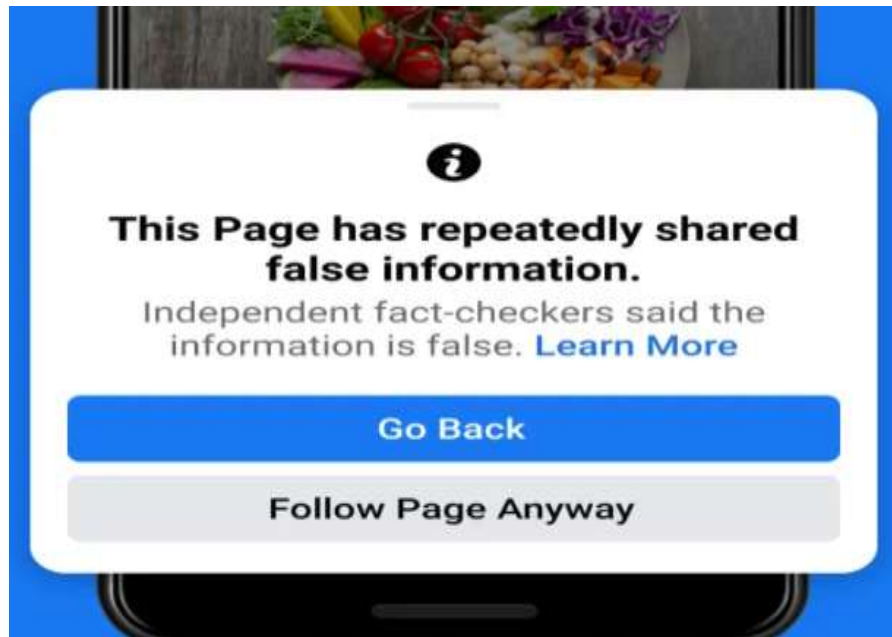
#### Contribution:

Easy to use, suitable for coursework and academic projects.

#### Limitation:

Mostly English-only articles.

#### 4. Facebook's Automated Fact-Checking



*Figure 4: Facebook's Automated Fast-Checking*

Facebook uses ML models to detect suspicious posts and then sends them to third-party fact-checkers. The system:

- Reduces the distribution of suspected fake posts.
- Adds caution labels
- Littered with links to verified sources



## 5. Google's Fact-Check Tools



Figure 5: Google's Fact-Check Tools

Google Search highlights fact-checked articles and provides context around potentially misleading claims.

## 6. Twitter/X Community Notes

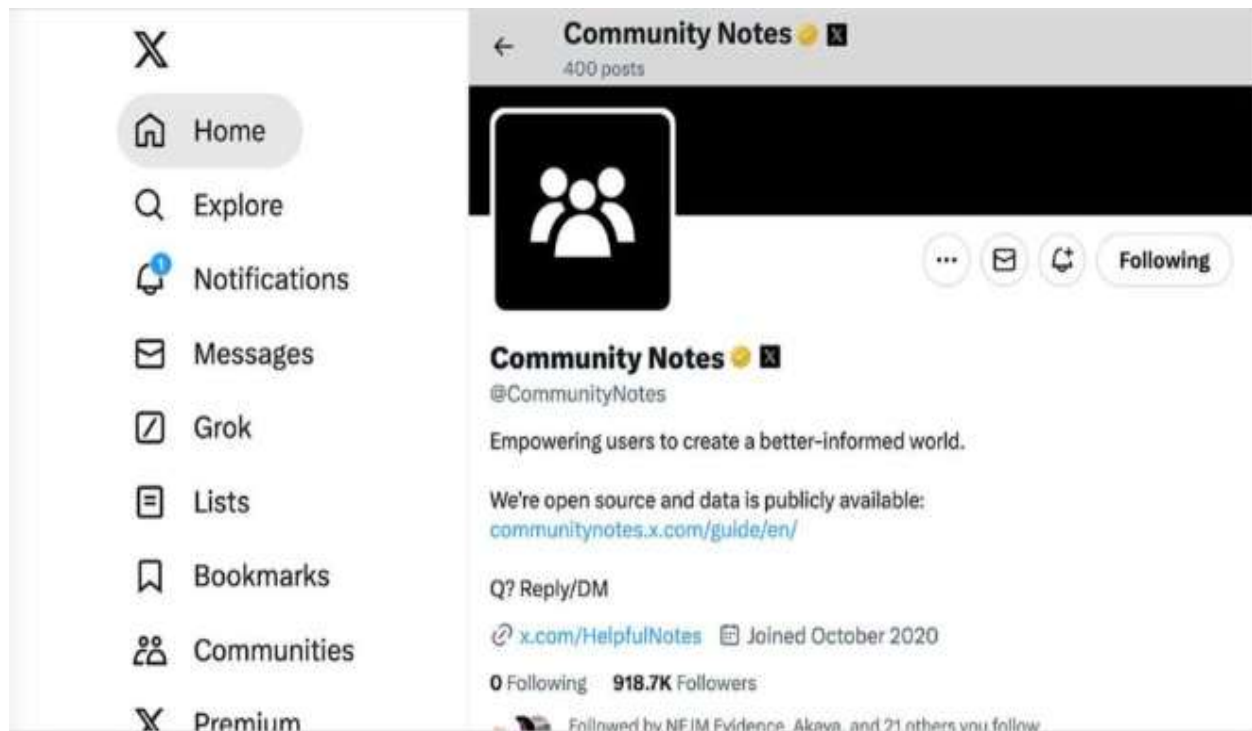


Figure 6: Twitter/X Community Notes

Twitter/X allows users to add context to misleading posts. Although not ML-based, it is a crowd-sourced misinformation control mechanism.

## 7. NewsGuard



*Figure 7: NewsGuard*

NewsGuard ranks news websites on their credibility, transparency, and reliability using manual and automated techniques. Analysis of Related Work Strengths The big volume of misinformation is quickly detected by existing systems. Because many of the datasets and models are publicly available, research has been easy for students and developers. Weaknesses: Fake news is an ever-evolving element, making it very hard to detect. Some models have difficulty identifying sarcasm, humour, or context. Misinformation in regional languages, like Nepali, is less studied. Large datasets and high computing power are required by deep learning models. Gap Identified: Current systems mostly target English-language datasets. Most don't incorporate social context or user engagement behaviour. That, in itself, provides a great opportunity for research using simpler ML/NLP models suitable for academic course work (UK-Parliament, 2020).

## **2. Background**

### **2.1. Research Work Done on the Fake News Problem**

Over the last decade has gained major attention in the field of Artificial Intelligence (AI) Natural Language Processing (NLP) and Machine Learning (ML). the rapid growth of social media platforms, online news portals, and digital communication has made it very easy for false or misleading information to spread quickly among people. Because manual verification of new article is impossible, researchers have focused on building automated systems that can identify fake news accurately and efficiently.

Research in this domain mainly focuses on three important areas: linguistic analysis of news text. machine learning based classification, and advanced deep learning approaches.

#### **1. Linguistic Pattern Analysis Using NLP**

Various studies have identified a fact that fake news, very regularly, follows some certain writing pattern that may not be so common when it comes to real news. These language characteristics mean identification by researchers for suspicious content. The common patterns of fake news include:

- Emotional and sensational language used to get attention.
- Clickbait headlines designed to mislead readers.
- Overuse of extreme terms like shocking, breaking or exposed.
- No trusted source of verifiable facts.
- Informal writing compared to professional journalism.

To study of these patterns based on their educational uses, calls for the use of various NLP techniques, including:

- Tokenization to break text into words.
- Stop word removal in order to get rid of common but meaningless words.

- lemmatization the word to their root form sentiment analysis to identify emotional tone.

These are the technique that help computers find the difference between fake and real news in terms of writing style and also the uses of language.

## **2. Fake News Classification using Machine Learning**

After the text is subjected to a Natural Language Processing (NLP) algorithm are used to classify the news is fake and real. A number of research paper established that the normal supervised machine learning algorithm is very efficient in carrying out the task of text classification along with the extraction techniques of TF-IDF.

Some of the machine learning algorithms commonly applied in studies of fake news are:

- Logistic Regression
- Naïve Bayes Classifier
- Random Forest
- Support Vector Machine
- Decision Tree

Algorithm derived from the kind of techniques are trained on data that has been reclassified as news and then predict which category a given story be belongs to base on its characteristics.

Various algorithms are also compared by researchers on various metrics such as accuracy and robustness by the algorithm for classifying news as a specific category.

### **3. Deep Learning Approaches**

As AI technology evolved, deep learning became popular for fake news detection:

- LSTM - Long Short-Term Memory networks for sequence learning.
- Text Classification with Convolutional Neural Networks (CNN)
- Transformers and BERT-based models of semantic understanding.

These models provide much higher accuracy because they understand context, meaning, and relationships between words better than traditional ML models. Overall, research in this domain is driven by the need for fast, accurate, automated systems capable of analysing millions of posts in real time.

## **2.2. Review and Analysis of the Related Work**

There exist many detection systems for fake news, as well as various datasets that can be used for the automation of fake news detection. The following is a description of how some of the most popular systems and datasets work, with the focus on mechanism and solutions.

### **A) LIAR Dataset (2017)**

#### **Mechanism:**

- Provides a structured dataset suitable for binary classification.
- Preprocessing steps include tokenization, stop word removal, and TF-IDF vectorization.
- ML models (Logistic Regression, Random Forest, SVM) are trained on features extracted from articles.
- Performance is evaluated using metrics like accuracy, precision, and recall.

#### **Solution Provided**

- Allows for multi-class classification of statements.
- Determines a fine level of truthfulness beyond fake or real.

## **B) Fake Newsnet Dataset**

### **Mechanism:**

- Integrates text analysis with features such as user interactions and posting behaviours.
- Content-based features (keywords, linguistic cues) are joined together with behavioural features (engagement, propagation).
- Graph methods and techniques of network analysis trace the spread of disinformation on social networks.

### **Solutions:**

- Provides support for advanced detection, which incorporates text and user behaviour.
- Assists in recognizing and detecting misleading and fake news through content and its spread on social platforms.

## **C) Kaggle Fake News Dataset**

### **Mechanism:**

- Offers a well-structured dataset for binary classification.
- The preprocessing steps include tokenization, removal of stop words, and TF-IDF transformation.
- Models based on Machine Learning (Logistic Regression, Random Forest, and SVM) are trained on article-level feature extractions.
- It is measured using factors such as accuracy, precision, and recall.

### **Solutions Provided:**

- Easy-to-use dataset for coursework and research purposes.
- Enables the execution of trials on varied machine learning algorithms and preprocessing processes.

## **H) Facebook's Automated Fact-Checking**

### **Mechanism:**

- ML algorithms processed content for linguistic patterns, keywords, and metadata.
- Flagged posts will be screened for authenticity from authorized sources.
- "False posts" deemed unreliable have "caution" notices placed alongside them, as well as links to trustworthy sources for verification.

### **Solutions Provided:**

- Prevents spread of suspected fake news.
- Supplies the user with validated context and alerts.

## **4. Google's Fact-Check Tools**

### **Mechanism:**

- Aggregates information gathered by fact-checking agencies and sites.
- Automatic matching of content to fact-checks.
- Shows search results containing context, rankings, and links to trusted sources.

### **Solutions Provided:**

- Assists users in making a distinction between accurate and inaccurate posts.
- It makes online information transparent and credible.



## **5. Twitter/X Community Notes**

### **Mechanism:**

- Platform where the community adds annotations to tweets with explanations.
- The ML model could point out posts requiring attention.
- "Community-approved notes" are also available to be viewed along with the tweet.

### **Solutions Provided:**

- It makes possible real-time misinformation spreading control.
- Raises awareness of deceptive content before the content goes viral.

## **6. NewsGuard**

### **Mechanism:**

- Incorporates manual editorial judgments with computer-driven checks for website activities.
- Asks about the level of transparency of ownership, accuracy of reporting, and journalistic standards.
- Provides users with the benefit of assigning credibility scores and explanations.

### **Solutions Provided:**

- Helps to navigate to authentic news sources.
- Minimizes the risk of exposure to incorrect information by pointing to credible sources. With this new version, the emphasis is on mechanisms and solutions rather than a series of contributions and limitations. This is a perfect fit for your teacher's request.

### **3. Solutions**

#### **3.1. Explanation of the proposed solution to solving the problem**

Basically, in this project I am building a system that is capable of automatically detecting the news article being fake or real. My goal is to build such a solution which can help people identify misleading or harmful information on social media platforms or online news. It is because fake news spreads quickly among many people, causing public confusion. Thus, an AI system that could review news articles and classify them would be very useful in real world.

The idea behind my solution is rather simple:

I use past examples to train a computer to understand the difference between fake and real news. Once the computer launches this pattern it can analyze any news text and decide whether it looks trustworthy or suspicious.

**This my system performs the following step:**

##### **A. Data Collection**

Basically, the dataset contains a large number of real and fake news but all of them have already been labeled as either fake or real to allow the machine to learn.

##### **B. Cleaning and preparing the text**

Text contains much noise in the form of punctuation special characters repeated words and common words like the and is these do not help in detecting fake news so the system removes them this cleaning process makes the text more meaningful.

##### **C. Features Extraction Using TF-IDF**

After cleaning, the system has to convert text into numbers because machine learning models cannot understand words directly. I make use of an NLP method known as TF-IDF (Term Frequency – Inverse Document Frequency), which gives importance to words

based on how frequent they are in a news article and how unique they are across the article.

This helps the system focus on meaningful words that contribute to identifying fake news.

#### **D. Training machine learning models**

The numerical data is split into training and testing sets. Multiple machine learning algorithms are trained using the training data:

- Logistic Regression
- Multinomial Naïve Bayes
- Support Vector Machine
- Decision Tree
- Random Forest

Each model learns different patterns from the same data.

#### **E. Model Evaluation and Comparison**

After training, I test the models on new articles that they have never seen. I do this so I can check how well they will be able to classify news between real and fake. Then, I compare the accuracy of both models and choose the one that has the better result.

#### **F. Predicting of News Articles**

The selected best model will be further used in predicting the new text entered by the user whether it is fake or real. This approach is not only easily implementable but also representative of how real-world fake news detection systems work.

These step by step is not only easy to implement but also reflects how real-world fake news detection system work.

### **3.2. Explanation of the AI algorithms used**

My solution uses NLP and ML. these two fields cooperate with each other: NLP deals with text and the prediction part is handled by ML. Each aspect is explained in one more detail below.

#### **A. Natural Language Processing (NLP)**

Natural Language Processing is an Artificial Intelligence-based science that deals with enabling computers to understand, interpret, and work with human languages in a meaningful manner. It helps the machine study, analyze, and draw sense out of texts or speech in the same way as human beings do. NLP teaches everything that comes under the purview of language: words, grammar, tone, and even hidden meanings.

NLP is a crucial because human language is messy and full of variations, like slang, spelling mistakes, short forms, emotions, sarcasm and various writing styles. These computers are not capable of understanding by themselves, So NLP comes into play with the techniques and rules to break up the languages into smaller part, study it, and convert it into something computer understand (Stryker, n.d.).

**In the task related to fake news detection, NLP plays an important role because:**

- The news is written in text, which NLP will be able to analyze.
- With NLP, the system will extract keywords, understand the meaning of every sentence, and find out writing style patterns.
- It does this in preparing the text such that machine learning correctly classifies it as either real or fake.

## **NLP techniques used:**

### **1. Cleaning Text**

This step removes unnecessary items like punctuations, special characters, extra spaces, numbers, meaningless words which make text simple in structure after cleaning and easier to analyze.

### **2. Tokenization**

This means breaking sentences into individual words.

For instance: "COVID-19 cases rise rapidly" → ["COVID-19", "cases", "rise", "rapidly"]

### **3. Stemming / Lemmatization**

These change words into their root form:

running → run

studies → study

This helps the model understand similar words as one.

### **4. TF-IDF Vectorization**

TF-IDF transforms every news article into a numerical vector of the importance of each word within it. The objective will be to give higher weights to words that appear more in fake news.

#### **TF-IDF is powerful because:**

- It reduces noise.
- It identifies meaningful words
- It handles large texts with ease.

## **B. Machine Learning (ML)**

It is a branch of Artificial Intelligence that enables computers to learn from data without being directly programmed to perform every single step. As opposed to telling a computer rules, we tend to provide it with a multimedia of examples, the machine then studies those examples seeking patterns, relationship and difference. Machine learning focuses on teaching a computer to make better predictions by learning over time, which is pretty similar to the way humans learn from practice (Chen, 2024).

### **Machine learning becomes pivotal in detecting fake news because:**

- It basically learns the difference between real and fake news through the study of numerical features created from news articles.
- This will help the system understand and recognize hidden patterns in writing style, word choice, and sentence structure.
- First, Machine Learning may be trained on numerous examples, after which it can take a new article and predict whether it is real or fake.
- Machine learning plays the role of decision maker in the project by helping the system classify news items with a good deal of accuracy and consistency.

In this project we use several supervised machine learning algorithms to detect fake news. Each algorithm has its own way of analyzing text and making predictions. Here's a breakdown:

### **1) Logistic Regression**

Logistic Regression is a simple and commonly used algorithm for classification problems, especially when the output is binary, like fake or real news. It predicts the probability that an article belongs to a certain category by analyzing the relationship between words in the text and labels. Logistic Regression is easy to interpret, fast to train, and works well even with large set of text features, such as TF-IDF.

#### **Role in the project:**

- Checks how words are related to fake or real labels.
- Predict the probability of a new article being fake or real.
- Helps find linear patterns in the text data.

#### **Why we use it:**

- Simple and efficient for text classification.
- Acts as a baseline to compare with more complex models.

## **2) Multinomial Naïve Bayes**

Multinomial Naïve Bayes is a probability-based algorithm often used in text classification task like spam detection and sentiment analysis. It calculates the probability of an article being in a certain class based on how often words appear. Even though it makes simple assumption, it often performs very well for text data and handles many features easily.

### **Role in the project:**

- Use word frequency to determine if a news article is fake or real.
- Provides a fast and reliable baseline for classification.

### **Why we use it:**

- Works well with TF-IDF features.
- Easy and quick to train on large datasets.
- Captures probabilistic patterns in word usage.

## **3) Support Vector Machine (SVM)**

SVM is a strong classification algorithm that separates data by finding the base boundary between classes. It works well in high-dimensional spaces, which is typical for text data after TF-IDF vectorization.

### **Role in the project:**

- Separate fake and real news by maximizing the margin between the two classes.
- Detects subtle difference in text patterns that other models might miss.

### **Why we use it:**

- Performs well with many features.
- Produce reliable results for complex data.
- Complements other models by focusing on margin-based decision rules.



#### **4) Decision Tree**

Decision Tree is a rule-based algorithm that splits data into branches based on feature values. It creates a tree structure where each node is a features or word, and each branch is a decision path leading to a prediction.

##### **Role in the project:**

- Classifies news using simple if-then rules.
- Provides interpretable results, showing which words or pattern influence predictions.

##### **Why we use it:**

- Easy to understand and explain.
- Captures non-linear patterns in the data.
- Can handle both categorical and numerical features from text.

#### **5) Random Forest**

Random Forest is an ensemble method that builds many decision trees and combines their predictions. By averaging or voting across trees, It reduces overfitting and increase accuracy.

##### **Role in the project:**

- Use multiple decision trees to classify news articles as fake or real.
- Reduce errors that a single decision tree might make.
- Produce more stable and consistent predictions.

##### **Why we use it:**

- Works efficiently with high-dimensional text features.
- Provide robustness and better generalization.
- Allow comparison with similar models like Logistic Regression and Naïve Bayes.

## **Why using Multiple Algorithm is important**

Using different algorithms helps create a stronger and more reliable fake news detection system because each model looks at the data differently:

- Logistic Regression: Captures linear relationships between words and labels.
- Multinomial Naive Bayes: Focuses on probabilities of word occurrence.
- SVM: Finds optimal boundaries in the high-dimensional space.
- Decision Tree: User rule-based paths for predictions.
- Random Forest: Combines many trees to reduce errors and improve stability.

By training and comparing all these models, we can find the best-performing one or even combine the result for better accuracy. This approach ensures more reliable predictions and demonstrates a good understanding of machine learning techniques for text classification.

### **3.3. Pseudocode of the Solution**

Pseudocode describes in detail, step by step, how a program or algorithm works. It is used to describe in simple, everyday English, how something works instead of using the real code. In other words, it is a descriptive method of representing the complete logic of a task that could be read by anyone, whether he or she knows coding or not. Pseudocode does not follow strict syntax rules like Python, Java, or C++. Instead, it expresses an idea, flow, and structure of the solution in the simplest way possible.

It acts as a bridge between thinking and coding. First, you write pseudocode to organize your thoughts, plan the logic, identify errors early, and make sure that everything makes sense. Then, you can take that pseudocode and turn into actual code. Pseudocode is language-independent, and it helps students, beginners, and even professional developers clearly communicate algorithms, explain processes, and design programs before the actual writing of the code.

#### **Why pseudocode is used in this project:**

- This makes the workflow of your code much easier to understand.
- It helps to explain the logic behind the fake news detection system.
- It exhibits a step-by-step procedure.
- Useful for planning the code before writing the real program.
- Helps to explain the process of non-technical readers.

### **3.3.1. Pseudocode of Fake News Detection System**

BEGIN

1. Load dataset containing labeled news articles (Fake and Real)
2. Preprocess the text data:
  - 2.1 Convert all text to lowercase
  - 2.2 Remove punctuation, numbers, and special characters
  - 2.3 Remove stop words (common unimportant words)
  - 2.4 Apply restricting or lemmatization to reduce words to their root form
3. Convert cleaned text into numerical features using TF-IDF vectorization
4. Split the dataset into:
  - Training set (80%)
  - Testing set (20%)
5. Select machine learning models for training:
  - Logistic Regression
  - Multinomial Naive Bayes
  - Support Vector Machine (SVM)
  - Decision Tree
  - Random Forest
6. Train all selected models using the training dataset
7. Test all models using the testing dataset
8. Evaluate performance of each model:
  - 8.1 Calculate accuracy, precision, recall, and F1-score

8.2 Compare performance metrics to select the best-performing model

9. Save the selected best model for final predictions

10. Input a new news article from the user

11. Preprocess the new article using the same cleaning steps as above

12. Convert the new article into TF-IDF features

13. Use the trained model to predict the class:

IF probability of "Fake News" > threshold:

    RETURN "Fake News"

ELSE:

    RETURN "Real News"

END

### 3.4. Diagrammatical representations of the solution

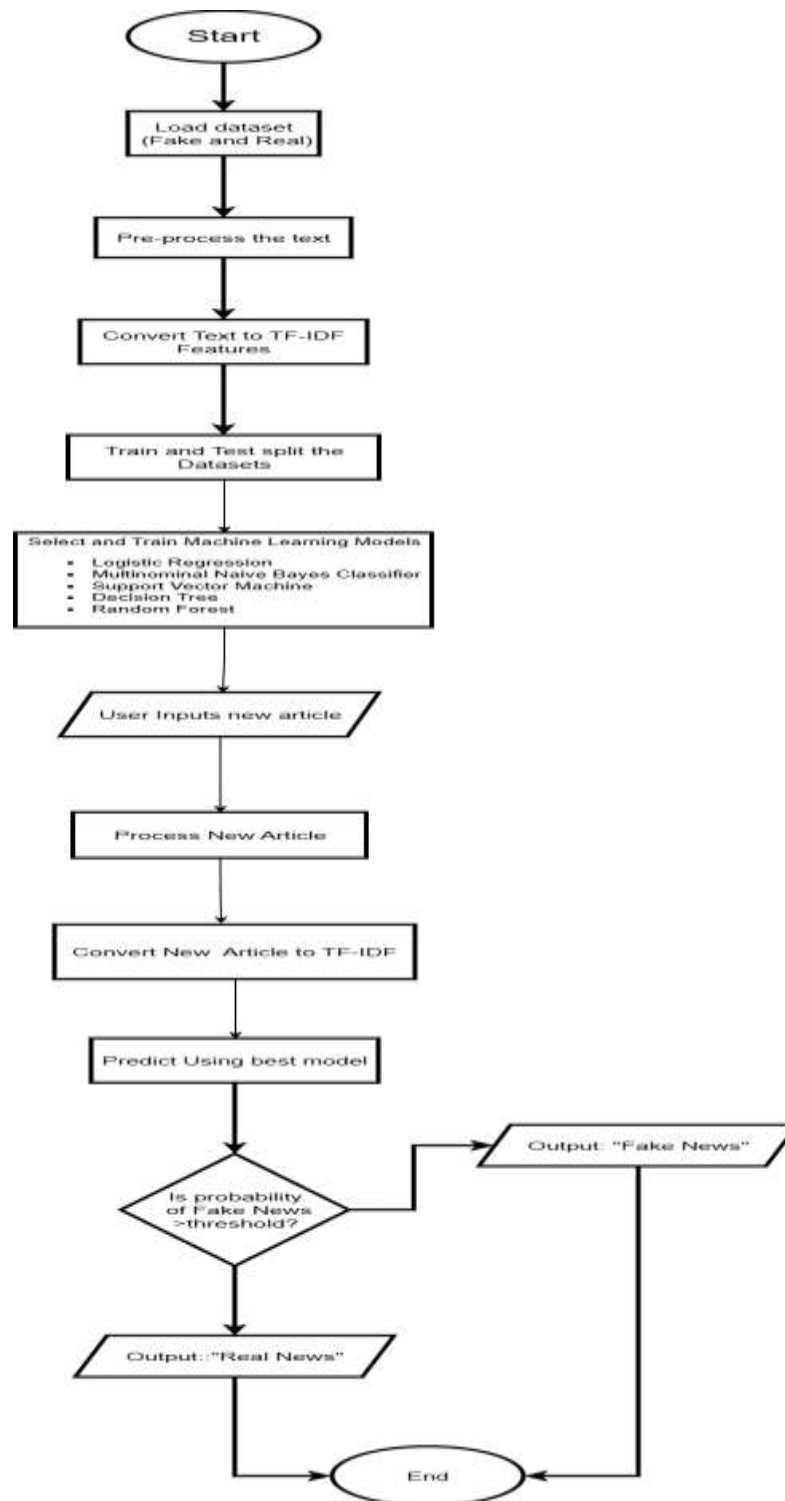


Figure 8: Flowchart

## 4. Background

### 4.1. Analysis of the work data

We created and developed a Fake News Detection based on Machine Learning techniques. In this project, TF-IDF features were used to train multiple machine learning algorithms, including Logistic Regression, Multinomial Naive Bayes, Support Vector Machine (SVM), Decision Tree and Random Forest. Each model was evaluated to understand its strength and weakness in classifying fake and real news. This comparative analysis helped identify the most suitable algorithm for the task. Our prime objective here is to develop an automated system that will be capable of distinguishing actual news and fake news based on the words present within it.

Throughout all these steps, we have followed a structure approach:

- A set of news articles with actual and false information was gathered.
- To clean the text, we removed punctuation, made lowercase, removed stopwords and normalized the text.
- We employed TF-IDF to transform the raw text data into significant numbers that our model could interpret.
- We have trained a Train Naive Bayes classifier. It identifies how often false and true news differ.

We then attempted to check the accuracy with which it could classify new, unseen news articles. However, on the whole, it was successful at demonstrating Machine Learning capabilities and abilities within text and making intelligent predictions. The system itself functions properly because TF-IDF and other models are excellent tools for text classification problems.

## **4.2. How the Solution Works for Real-World Situations**

The issues of fake news have become prominent within today's online world. Many people rely on online source information, and misleading information can be propagated on online platforms. The solution directly helps reduce the impact of fake news.

**The solution addresses real-world problems in these ways:**

- **Assists in identifying deceptive or damaging information**

An individual can rely on this system to check very swiftly if a news article might be a replica or not before accepting it.

- **Helps fact-checking bodies**

It will be useful for news-verification organizations as they will be capable of making use of ML tools that will enable them to automatically detect news.

- **Reduces spreading of misinformation on social platforms**

Platforms can incorporate models for identifying fake news so that they can either prevent or alert users regarding suspicious news.

- **Enhances media literacy**

By learning about ways in which 'fake news' can be detected, it leads people to be more aware of 'misinformation'.

- **Offers fast and scalable detection**

Opposing to human checking, ML technology can analyse thousands of articles in simple seconds. The model for this project is not flawless, but it does offer a great platform for the role that technology can play within journalism and the safeguarding of society against misinformation.



### **4.3. Further Works**

The project works well there still needs to be improvement the world of fake news developed constantly and the system needs to evolve as well.

**Future developments might include:**

- **Using Deep Learning Models like LSTM or BERT**

These models are better at understanding meaning and can make more accurate predictions.

- **Increasing the amount of data**

A larger and more varied set would assist the model in identifying more varieties of misleading news.

- **Adding Image/Video Analysis**

Often, there are manipulated images and videos. A multimodal model would be capable of handling text as well as images.

- **Identifying the emotive tone associated with fake news**

"Stand with 'Fake news' may employ use of emotive words (fear, anger). Sentiment analysis would make predictions better.

- **Real-time detection tool**

A browser extension or app could enable an immediate check of news.

- **Creating an appropriate web interface**

Creating an appropriate Easy-to-use UI for news entry and getting immediate results.

- **Improving interpretability**

Adding explanations such as: “This news seems to be pseudo because it contains these words.”

- **Applying ensemble techniques**

Using multiple ML models to improve prediction accuracy.

## References

- Chen, M. (2024, November 25). *What Is Machine Learning?* Retrieved from Oracle:  
<https://www.oracle.com/in/artificial-intelligence/machine-learning/what-is-machine-learning/>
- Mao, Z. (2025). An overview of fake news detection: From a new perspective. *Fundamental Research*, 332-346. Retrieved from ScienceDirect:  
<https://www.sciencedirect.com/science/article/pii/S2667325824000414>
- Sastrawan, K. (2021, September 15). *Fake News Dataset*. Retrieved from Mendeley Data: <https://data.mendeley.com/datasets/945z9xkc8d/1>
- Stryker, C. (n.d.). *What is NLP (natural language processing)?* Retrieved from IBM:  
<https://www.ibm.com/think/topics/natural-language-processing>
- UK-Parliament. (2020). *Response to the House of Lords Communications and Digital Committee's*. UK: UK-Parliament.
- Wang, W. Y. (2017, 05 01). *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*. Retrieved from Emergent Mind:  
<https://www.emergentmind.com/papers/1705.00648>