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. Explanation of the Topic and AI Concepts Used

Artificial Intelligence (AI) can be defined as a field of computer science that aims at creating systems which can perform tasks requiring human intelligence. Human intelligence includes learning from experiences, language understanding, pattern recognition, and decision-making.

The project uses Artificial Intelligence in terms of Machine Learning (ML) and Natural Language Processing (NLP), which is used to solve the issue of detecting fake news.

A. Machine Learning (ML)

Machine Learning enables a computer system or model to learn automatically from data as it is not necessarily predefined or programmed to accomplish every task. In supervised machine learning, the model is trained on labeled data so that is able to learn the relationship between the input and output (Chen, 2024)..

- The input for this project is the text of the news articles.
- The output in this case, is a label that specifies whether the news is ‘Fake’ or ‘Real’.
- The system relies on past, labeled examples to develop the ability to categorize new, unseen news articles.

Machine Learning is appropriate for this purpose because:

- Fake News has its own typical patterns concerning the writing style and words used.
- The patterns can be effectively learned by ML algorithm.
- ML algorithms are capable of handling high volumes of data quickly and accurately.

B. Natural Language Processing (NLP)

Natural language is an area of AI that allows computers to read, process, and analyze human language. Because news article contains human language, Natural Language Processing becomes an essential part of this initiative as well (Stryker, n.d.).

NLP is required because:

- Computers do not interpret human language.
- The text data is unstructured and noisy.
- Text is thus turned into a mathematical representation that can be processed by ML models.

NLP is employed in this project to:

- Clean Raw Text
- Extract significant words.
- Represent text as numeric features using TF-IDF transformations.

NLP works together with ML for the classification of the data, hence the effectiveness of the system for the detection of the fake news

1.2. Explanation of the Chosen Problem Domain

Fake News can be described as information that is incorrect or misleading, but it appears as real information. In the current digital era, fake news has been spreading faster through social media platforms, online new sites, and messaging applications. People commonly share information without verifying it which has been causing this rise of misinformation.

Problems Caused by Fake News

- Fake news creates several serious problems in society:
- It misleads people by providing incorrect information.
- It can influence public opinions and political decisions.
- It may create fear, panic, or social conflict.
- It reduces trust in reliable and verified news sources.

Why Fake News Is Difficult to Control

- A large amount of content is shared daily on the internet.
- News spreads faster than physical confirmation.
- Manual fact-checking is slow and costly.
- It’s not possible for a social to check each piece of content on the Internet.

Need for an AI-Based Solution

- An AI-based solution is required to manage fake news effectively:
- AI can process large datasets quickly.
- Machine learning models can identify patterns in fake news.
- Automated systems provide fast and scalable detection.

Objective of the Project

The objective of this project is to develop a machine learning–based fake news detection system that automatically classifies news articles as Fake or Real, helping to reduce the spread of misinformation and improve trust in digital news.

2. Background

This section explains the background study completed in my Proposal. It summarizes existing research related to fake news detection and provides a base for developing the proposed system.

2.1. Research Work on Fake News Detection

Detection of fake news has become a popular area of research in the last few years as the growth of online media has increased rapidly. Scientists and researchers from the fields of AI and Natural Language Processing (NLP) have proposed various solutions for the detection of fake news using various methods.

The majority of the work done on this research area focuses on three domains:

- Language-based identification, in which the writing style in news articles is analyzed. Often, the writing in what is labeled as fake news is full of encouragement language, is interesting in titles, and depends on unauthorized sources.
- Machine learning algorithms where the text of news stories is represented in numeric form and classifiers learn to distinguish between false and true news.
- Deep Learning models, such as LSTM, CNN, and BERT, which are capable of detecting complex text patterns but require more substantial amounts of data and heavier computing capabilities.

In the case of coursework projects, conventional machine learning techniques are often adopted as they are relatively easy to implement, easy to interpret, and still guarantee relatively accurate results.

2.2. Research Work on Justification

For this project, the fake_new_dataset_csv is selected. This dataset contains a large collection of news articles, each clearly labeled as fake or real, making it suitable for binary classification.

The dataset is chosen for the following reasons:

- It is frequently employed in research work.
- It is capable of supervised machine learning.
- The English language is used for the data, making it easy to pre-process.

3. Solution

This topic explains the solution proposed to address the problem of fake news detection using Artificial Intelligence techniques. It explains the algorithm used, the working procedure, the development tools, as well as the results obtained.

3.1. Explanation of the Solution and Used AI Algorithm

This problem will be solved by implementing an approach based on machine learning, which will decide whether a particular news post on this website is Fake or Real. This approach will use the content of news articles to decide

In the current project, the machine learning algorithm implemented is of the supervised type. The model is trained on labeled data samples where the news article is already classified as being false or probably true. The model relies on the differences in the word choice and writing styles.

The approach involves these broad steps:

1. Data Collection

A data set is well known having news articles considered as either Fake or Real to train the model. Every news article contains of written data and its label.

2. Text Processing

- The process removes unwanted characters like punctuation, numbers, and special symbols.
- All text will be in lower case letters.
- The words such as the, and is are deleted as they don’t help in the process of classification.
- The words are reduced to their origin through lemmatization or stopping. For example, runs-run, studies-study.

3. Feature Extraction with TF-IDF

- Text data is changed into numerical from because machine learning algorithms are not able to process words straight.
- TF-IDF (Term Frequency Inverse Document Frequency). This is used for heavy task based on importance.
- The words that are commonly observed in the fake news but not in the real news have a high importance value, which helps to identify patterns.

4. Machine Learning Algorithms Employed

Several supervised machine learning algorithms are used to classify the news:

- Logistic Regression

It is used to predicts the probability of an article being fake or real using relationships between words and labels.

- Multinomial Naïve Bayes

This classifier makes use of word frequency probability to identify the class into which a news article belongs.

- Support Vector Machine (SVM)

It separates both fake and real news and use the suitable border on high-dimensional data.

- Decision Tree

Classifies news articles using the if-then rule of word usage in terms of a tree format.

- Random Forest

The outputs from different decision trees are combined for more correct results.

5. Model Training and Testing

- Every dataset is divided into a training set and a testing set. In this case, the dataset is divided into an 80% training set.
- Models for each choice are trained on the training data and tested using the test data.
- The model having the best performance will be chosen regarding prediction on new articles.

3.2. Pseudocode of the Solution

Pseudocode is the step-by-step process that explains in detail the way a program or algorithm works. Pseudocode is applied to describe the way something works in a normal and understandable way by using English rather than the actual programming language. This is to say that pseudocode is a way of understanding the full logic of a given task in a way that can easily be read and comprehended by a person even if he does not know coding. Pseudocode is not bound by specific syntax rules such as in the case of Python, Java, and C++ languages.

It is a bridge between thinking and coding. Firstly, a person has to write a pseudocode to have their thoughts organized, understand the logic, spot mistakes, and ensure that all is well. After this is done, one can easily translate this pseudocode to actual coding. A pseudocode is language independent, and it assists students, new coders, as well as professionals, to ensure that algorithms are communicated, processes are described, and programs are designed even before any coding takes place.

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.

3.2.1. 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

3.3. Diagrammatical representations of the solution

3.4. Explanation of the development process

In this project, I created Fake News Detection system using Python programming along with Machine Learning techniques. In this system, I applied Natural Language Processing (NLP) concepts to classify news as Fake or Real news. This project is created with the help of Anaconda Navigator, which is a full-fledged Python distribution solution for package management. This further made it simpler to execute this project in Jupyter Notebook.

3.4.1. Tools Used

- Anaconda Navigator

It is a tool that helped me manage Python, Libraries, and environments. I used it to launch a Jupyter Notebook and Install useful Python libraries. It made my work smooth and easier because I didn't have to use command lines all the time. It ensures smooth running of the application without conflicts between libraries.

- Jupyter Notebook

It is where I did most of my coding work. I used it to write Python code for data cleaning, analysis and creating visualizations like graphs and charts. It was helpful because I could see the output of my code just below each cell, which made it easier to understand and fix mistakes.

- Python Language

Python was used for its simplicity and strong support for AL and machine leaving. It works well with libraries nltk and scikit-learn, and runs smoothly in Anaconda Navigator for easy coding and testing.

3.4.2. Libraries and Dataset Used

For the project, Python libraries were employes for text processing, model training, and outcome analysis.

PANDAS: For the efficient handling of the data set. It was used for reading the Excel file, retrieving the needed columns (text and label), and dropping the missing observations.

RE: For removing text cleaning operations like removing stopwords and taking the base form of the words punctuations numbers and special characters.

NLTK: For Natural Language Processing as in removing stopwords and taking the base form of the words.

SCIKIT: LEARN: Used for text to TF- IDF transformation, splitting the dataset, training classifier (Logistic Regression, Naïve Bayes, SVM) calculating accuracy.

In the fake_news_data.csv file, the text data is in the text column, and the labels are in that label column (0 represents Fake, 1 represents Real). The tools enabled efficient preprocessing, feature selection, and classification of the news article.

3.4.3. Stepwise Explanation of the Code

Step 1

First, we enhanced necessary Python libraries to handle the dataset pandas, cleaning and processing text with re and nltk, feature mining and training machine learning models with scikit-learn, and performance calculation.

Step 2

At this stage, we need the dataset “fake_new_dataset.csv” using the pandas library. We choose to read only the text column and the label column from the dataset because missing values affect the dataset: therefore, we eliminate the missing ones to ensure the data set is ready for processing.

Step 3

This step should describe the cleaning of text data in order to prepare it for analysis. Using NLTK, stopwords were removed, meaning common words such as “the” or “is”, and words were lemmatized to their base form, such a changing “running” to “run” to make the text consistent for the model.

Step 4

In this stage, the processed text data is converted to numerical features using TF-IDF. TF-IDF is used to calculate the weighting of words in each news article so that the models can interpret the data.

Step 5

The data is divided into a training set and a test set, with 80% of data in the training set 20% in the test set. The training set is used for training a model, while the test set is used for testing how well the model has performed on a new data set.

Step 6

In this step, three machine learning algorithms were trained on the training data. The three algorithms used were Logistic Regression, Multinominal Naive Bayes, and Support Vector Machine (SVM). These algorithms learn the patterns in the news data to determine whether the article was Fake or Real.

Step 7

Accuracy of the trained model for the test dataset is estimated. This is an indication of how effectively the model is able to classify the news as Fake or Real. For the project, the model with the best accuracy is generally the SVM model.

Step 8

A function was developed for predicting new news articles. The function removes unnecessary characters from the input news, and then TF-IDF vectors are obtained from the news and the trained SVM model classifies the views as Fake or Real.

Step 9

The program allows the user to provide news text. After that it relies on the prediction function for classification and shows Fake or Real news in real-time.

3.4. Achieved Results

In the screenshot, a news article considered as false news is selected from my dataset. This text is used to train the model to recognize patterns of false information and help the system differentiate fake news from real news.

After choosing the fake news article from the dataset, the text was entered into the system's input. The trained model processes the text and correctly classified it as Fake News, representing that the system can recognize false information correctly.

A news article labeled as true is selected from the dataset. This text is used to teach the model how real news looks so it can tell the change between real and fake news.

After selecting the real news article, the text was entered into the system. The model checks the text and correctly classifies as Real News, display that the system can know real news correctly.

When the input box is left blank and no news text is entered, the system shows "No News Entered". This make sure that the program manages empty input properly and does not try to process nothing.

4. Conclusion

In this milestone, a Fake News Detection System was designed and developed with the help of Python programming and machine learning. The dataset was cleaned and preprocessed. The text features were extracted on the basis of TF-IDF vectorization. Three models were designed and tested. The Fake News Detection System can classify a news article into Fake or Real categories. The higher accuracy was achieved by the SVM model.