VELAMMAL
INSTITUTE OF TECHNOLOGY

Approved by AICTE - New Delhi
Affiliated to Anna University - Chennai
Accredited by NBA & NAAC

NAAN MUDHALVAN PROJECT(IBM)

IBM AI 101 ARTIFICIAL INTELLIGENCE-GROUP 1

Team name: Proj_224823_Team_1

Team members: SAI VIGNESH S (reg no 113321106079)

THIRUMARAN S (reg no 113321106106)

VIGNESH V (reg no 113321106115)

LINGESH V(reg no 113321106301)

SRINIVASA BHARATH S(reg 113321106094)

Problem Statement :

Design and develop an NLP-based system that can accurately identify and classify news articles or information as either "fake" or "real" by analyzing the textual content, with the primary goal of mitigating the spread of misinformation and promoting the dissemination of trustworthy information.

DESIGN STEPS:

Data Acquization:

1. Data Collection:
   The first step is to collect the data from various sources such as news websites, social media platforms, etc. and load it into a dataset.

2. Data Labeling:

In this step, label the data as real or fake based on the authenticity of the news article. By using crowdsourcing platforms or pre-trained models to label the data.

Data Preprocessing:

Data preprocessing is the initial step , where raw data is cleaned, organized, and transformed into a suitable format for further processing.

preprocessing steps :

- **Data Cleaning:**

1. Removal of HTML Contents
2. Removal of Punctuation Marks and Special Characters
3. Removal of Stopwords
4. Lemmatization
5. Perform it for all the examples

- **Data Transformation**: In this step, Data transform into a suitable format for analysis by performing operations such as normalization, scaling, or encoding.

- **Data Integration**: In this step, Data combine from multiple sources, such as databases and spreadsheets, into a single format.

Tokenization:

1. Tokenization is the process of breaking down a text into smaller units called tokens, which can be words, phrases, or sentences.
2. It is a crucial step in natural language processing (NLP) and is used in various NLP tasks such as sentiment analysis, machine translation, and named entity recognition

Vectorization :

•Vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which is used to find word predictions, word similarities/semantics.

Model Training:

Train the selected model on the preprocessed data using algorithms such as

- Naive Bayes
- Decision Trees
- Logistic Regression, or
- Support Vector Machines (SVM).
- Passive Aggressive algorithms .

MODEL EVALUATION:

•For model evaluation of Fake news detection using NLP,use metrics such as accuracy, precision, recall, and F1-score.

•The accuracy of a fake news detection model depends on various factors such as the quality of the dataset, the choice of machine learning algorithm, and the feature extraction techniques used

•By using the trained model ,it can evaluate the authenticity of news articles.