COMPUTATIONAL LINGUISTICS AND NATURAL LANGUAGE PROCESSING

ASSIGNMENT-1

NAME: ASMI GARG	**SAP ID:** 500107577	**ROLL NO:** R2142220457
COURSE: B.TECH CSE(AIML-NH)	**BATCH:** B4	**SEMESTER:** VI

DESCRIPTION OF THE DATASET

FAKE NEWS DETECTION

Purpose of the project

The Fake News Detection project aims to develop an AI-based system that automatically classifies news articles as real or fake using machine learning and natural language processing (NLP). By analyzing textual features, writing styles, and source credibility, the model helps identify misinformation and prevent its spread. The project involves curating a balanced dataset of real and fake news, extracting key linguistic patterns, and building a scalable detection system to promote accurate information dissemination.

Dataset Overview

This project involves two datasets:

- True.csv: Contains 9,999 real news articles.
- Fake.csv: Contains 10,019 fake news articles.

Each dataset consists of three columns:

- Headline: The title of the news article.
- Content: The body text of the news article.
- Source: The media outlet or origin of the news.

Data Collection and Preparation

True News Dataset (True.csv)

- Collected using web scraping with BeautifulSoup from verified media sources.
- The Mediastack API (http://api.mediastack.com/v1/news) was used to extract headlines, content, and source information.
- Contains 9,999 articles from 489 unique sources.
- Top source: "lulegacy" (2,260 articles).

Fake News Dataset (Fake.csv)

- Generated using AI-based text generation models and supplemented with datasets from Kaggle.
- Contains 10,019 articles from 117 sources.
- Top source: "left-news" (2,236 articles).