**ASSIGNMENT-5**

PROBLEM STATEMENT:CLASSIFYING FAKE NEWS

Forest fires and fake news spread at asimilar speed.Sometimes fake newses can be threatening to life and cause emergency situations.Thus it is important to prevent the misinformation spread across the platforms.Fake news can be identified on the following characterstics.

Language:News in the form of text and string.

Since the data is enormous it is not possible to analyse the every single text statement by a person.A technique based on natural language processing is applied to identify the news in realtime.

Natural Language Processing: Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence.

Steps have to be followed for the project:

1.Understand the Problem statement and ensure the solution to solve the problem

2.Collect the suitable dataset for the problem statement of the Ml project

3.Install the Necessary libraries:

Pip install numpy pandas sklearn

4.Install the below packages:

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.linear\_model import PassiveAggressiveClassifier

from sklearn.metrics import accuracy\_score, confusion\_matrix

from nltk.corpus import stopwords

from nltk.stem.porter import PorterrStemmer

from sklearn.naive\_baye import MultinomialNB

5.Preprocessing the data by cleaning the data and removing all the null values

6. In-text preprocess we are cleaning our text by steaming, lemmatization, remove stopwords, remove special symbols and numbers, etc. After cleaning the data we have to feed this text data into a vectorizer which will convert this text data into numerical features.

7. **Stop words:**words that occur too frequently and not considered informative

Examples :

{‘the’, ‘a’, ‘an’, ‘and’, ‘but’, ‘for’, ‘on’, ‘in’, ‘at’ …}

***Split the Data***

Splitting the data is the most essential step in machine learning. We train our model on the trainset and test our data on the testing set. We split our data in train and test using the train\_test\_split function from Scikit learn.

We split our 80% data for the training set and the remaining 20% data for the testing set

we’ll predict on the test set from the CountVectorizer and calculate the accuracy with accuracy\_score() from sklearn.metrics.

Confusion matrix:Basically this matrix show how many results are correctly predicted and not correctly predicted.

Output:156 are false positives and 290 are false negatives

We got an accuracy of 90.82% with this model. Finally, let’s print out a confusion matrix to gain insight into the number of false and true negatives and positives.

Team Members

D.Keerthi-S180532

G.lakshmi Neeraja-S180746

CSE section- 2E