**FAKE NEWS DETECTION USING NLP**

**LOADING AND PRE-PROCESSING DATASET**

import pandas as pd

import re

import nltk

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

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

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.preprocessing import LabelEncoder

dataset = pd.read\_csv("C:/Users/SMCE-BIGDATA/Documents/fake\_news\_detection.csv")

print(dataset.head())

print(dataset.info())

def preprocess\_text(text):

text = text.lower()

text = re.sub(r'[^a-zA-Z]', ' ', text)

text = re.sub(r'\s+', ' ', text)

words = nltk.word\_tokenize(text)

stop\_words = set(stopwords.words('english'))

words = [word for word in words if word not in stop\_words]

stemmer = PorterStemmer()

words = [stemmer.stem(word) for word in words]

return ' '.join(words)

dataset['text'] = dataset['text'].apply(preprocess\_text)

label\_encoder = LabelEncoder()

dataset['label'] = label\_encoder.fit\_transform(dataset['label'])

X = dataset['text']

y = dataset['label']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

vectorizer = CountVectorizer(max\_features=5000)

X\_train = vectorizer.fit\_transform(X\_train)

X\_test = vectorizer.transform(X\_test)

**OUTPUT :**

title ... Unnamed: 171

0 Donald Trump Sends Out Embarrassing New Year’... ... NaN

1 Drunk Bragging Trump Staffer Started Russian ... ... NaN

2 Sheriff David Clarke Becomes An Internet Joke... ... NaN

3 Trump Is So Obsessed He Even Has Obama’s Name... ... NaN

4 Pope Francis Just Called Out Donald Trump Dur... ... NaN

[5 rows x 172 columns]

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 23502 entries, 0 to 23501

Columns: 172 entries, title to Unnamed: 171

dtypes: object(172)

memory usage: 30.8+ MB

None