FAKE NEWS DETECTION USING NLP

PHASE 3

IMPORTING NECESSARY LIBRARIES

```
import pandas as pd
import re
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

READING FILES

- There were two datasets, one consisting of fabricated or false information and the other containing genuine or authentic data.

```
fake_data=pd.read_csv(r"C:\Users\navee\OneDrive\Documents\FAKE NEWS DETECTION\Fake.csv")
```

```
true_data=pd.read_csv(r"C:\Users\navee\OneDrive\Documents\FAKE NEWS DETECTION\True.csv")
```

TEXT CLEANING

- Text cleaning involves removing noise and irrelevant elements from the text data to make it more suitable for analysis.
- Two types are:

    Special characters removal

    HTML tags removal

REMOVING SPECIAL CAHARACTERS

- Special characters, such as punctuation marks, symbols, and non-alphanumeric characters, are often removed. These characters may not carry meaningful information and can interfere with analysis.

```
fake_data['no_sc_text']=fake_data['text'].apply(lambda x: re.sub(r'[^A-Za-z0-9\s]', '', x))
fake_data['no_sc_text'].head()
```

```
true_data['no_sc_text']=true_data['text'].apply(lambda x: re.sub(r'[^A-Za-z0-9\s]', '', x))
true_data['no_sc_text'].head()
```

REMOVING HTML TAGS

- When dealing with web data, HTML tags are often present in the text. These tags are removed to ensure that the text is in a consistent and human-readable format.

```
fake_data['no_html_text'] = fake_data['no_sc_text'].apply(lambda x: re.sub(r'<.*?>', '', x))
fake_data['no_html_text']
true_data['no_html_text'] = true_data['no_sc_text'].apply(lambda x: re.sub(r'<.*?>', '', x))
true_data['no_html_text']
```

TOKENIZATION

- Tokenization is the process of splitting the text into individual words or tokens
- Word Splitting: Text is split into individual words, making it easier to analyze the content at a more granular level.

```
fake_data['tokenise_text'] = fake_data['no_html_text'].apply(lambda x: word_tokenize(x))
```

```
true_data['tokenise_text'] = true_data['no_html_text'].apply(lambda x: word_tokenize(x))
```

STOPWORD REMOVAL

- Stopword removal involves eliminating common words, known as stopwords, from the text.
- Stopwords: Stopwords are frequently occurring words in a language, such as "and," "the," "is," "in," and "of." These words are considered to be of little value in many NLP tasks as they don't carry significant meaning.
- Removal Purpose: Removing stopwords helps reduce noise in the data and can make text analysis more focused on content words, which are often more informative.

```
stop_words = set(stopwords.words('english'))
fake_data['sw_text']  = fake_data['no_html_text'] .apply(lambda x: ' '.join([word for word in x.split() if word.lower() not in st
true_data['sw_text']  = true_data['no_html_text'] .apply(lambda x: ' '.join([word for word in x.split() if word.lower() not in st
```

FINAL DATASET

true_data

	title	text	subject	date	no_sc_text	no_html_text	tokenise_text	sw_text
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	WASHINGTON Reuters The head of a conservative...	WASHINGTON Reuters The head of a conservative...	[WASHINGTON, Reuters, The, head, of, a, conser...	WASHINGTON Reuters head conservative Republica...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	WASHINGTON Reuters Transgender people will be...	WASHINGTON Reuters Transgender people will be...	[WASHINGTON, Reuters, Transgender, people, wil...	WASHINGTON Reuters Transgender people allowed ...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	WASHINGTON Reuters The special counsel invest...	WASHINGTON Reuters The special counsel invest...	[WASHINGTON, Reuters, The, special, counsel, i...	WASHINGTON Reuters special counsel investigati...
3	FBI Russia probe helped by Australian	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	WASHINGTON Reuters Trump campaign adviser Geo...	WASHINGTON Reuters Trump campaign adviser Geo...	[WASHINGTON, Reuters, Trump, campaign, adviser...	WASHINGTON Reuters Trump campaign adviser Geor...

fake_data

	title	text	subject	date	no_sc_text	no_html_text	tokenise_text	sw_text
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	Donald Trump just couldn t wish all Americans ...	Donald Trump just couldn t wish all Americans ...	[Donald, Trump, just, couldn, t, wish, all, Am...	Donald Trump wish Americans Happy New Year lea...
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	House Intelligence Committee Chairman Devin Nu...	House Intelligence Committee Chairman Devin Nu...	[House, Intelligence, Committee, Chairman, Dev...	House Intelligence Committee Chairman Devin Nu...
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	On Friday it was revealed that former Milwauke...	On Friday it was revealed that former Milwauke...	[On, Friday, it, was, revealed, that, former, ...	Friday revealed former Milwaukee Sheriff David...
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	On Christmas day Donald Trump announced that h...	On Christmas day Donald Trump announced that h...	[On, Christmas, day, Donald, Trump, announced,...	Christmas day Donald Trump announced would bac...
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	Pope Francis used his annual Christmas Day mes...	Pope Francis used his annual Christmas Day mes...	[Pope, Francis, used, his, annual, Christmas, ...	Pope Francis used annual Christmas Day message...