TITLE: FAKE NEWS DETECTION USING NLP

PHASE 3 : DEVELOPMENT PART 1

SUBMITTED BY :B.SUJI

Mail ID :suji17062004@gmail.com

Loading and Pre-Processing the Dataset

**INTRODUTION :**

This research addresses the critical elements of fake news detection: preprocessing and dataset. Preprocessing techniques are pivotal for preparing textual data, while dataset construction is fundamental for model training and evaluation. This paper explores effective preprocessing methods and highlights the significance of a well-curated dataset in advancing the field of fake news detection.

**ABOUT THIS PHASE :**

In this phase we need to do loading and pre-processing the datasets. Here I explain about what are the process to do this phase.

**DATA PREPROCESSING:**

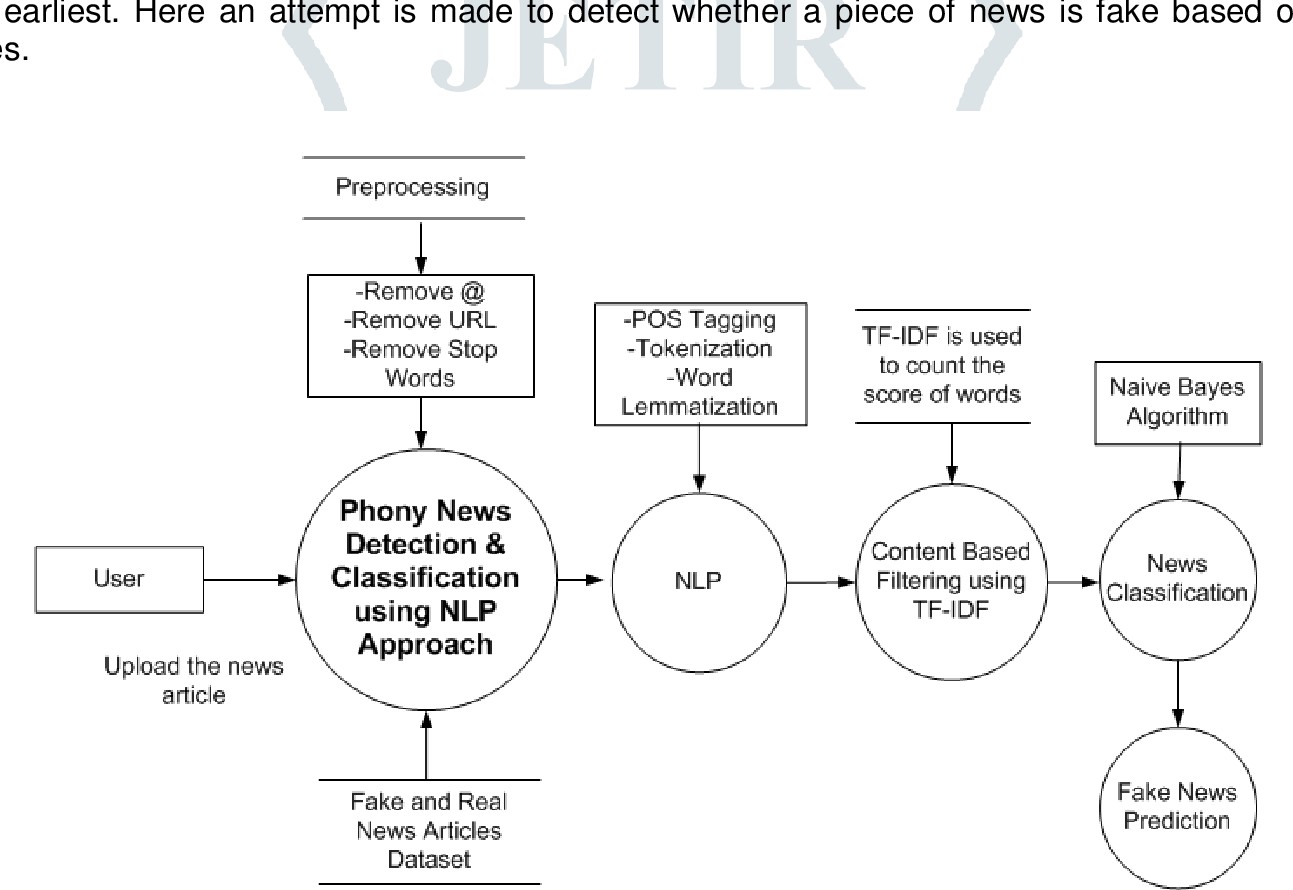
After gathering the data, you must preprocess it to make sure it is clean and suitable for analysis. This entails:

Handling missing data: If any of the fields have missing values, you may need to fill them up or delete the entries that correspond to them.

Data encoding: Genre and language categories have to be encoded into a category-specific format. This is achievable.utilizing methods like label encoding or one-hot encoding

Feature scaling: if you intend to employ algorithms that are sensitive to feature scales, such as gradient descent- based approaches, you might need to scale numerical features like runtime to have similar ranges.

Outlier detection: Find any unusual trends in the IMDb ratings or runtime that might skew the forecasts and deal with them.



**Importing Dataset:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

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

from sklearn.metrics import accuracy\_score

from sklearn.metrics import classification\_report

import re

import string

df\_fake = pd.read\_csv(“/content/Fake.csv.zip”)

df\_true = pd.read\_csv(“/content/True.csv.zip”)

df\_fake.head()

OUT:

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

df\_true.head(5)

Out:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | title | text | subject | date |
| 0 | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | politicsNews | December 31, 2017 |
| 1 | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | politicsNews | December 29, 2017 |
| 2 | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | politicsNews | December 31, 2017 |
| 3 | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | politicsNews | December 30, 2017 |
| 4 | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | politiesNews | December 29,2017 |

Inserting a column “class”as target feature

df\_fake[“class”] = 0

df\_true[“class”] = 1

df\_fake.shape, df\_true.shape

((23481, 5), (21417, 5))

# Removing last 10 rows for manual testing:

df\_fake\_manual\_testing = df\_fake.tail(10)

for i in range(23480,23470,-1):

df\_fake.drop([i], axis = 0, inplace = True)

df\_true\_manual\_testing = df\_true.tail(10)

for i in range(21416,21406,-1):

df\_true.drop([i], axis = 0, inplace = True)

df\_fake.shape, df\_true.shape

((23471, 5), (21407, 5))

df\_fake\_manual\_testing[“class”] = 0

df\_true\_manual\_testing[“class”] = 1

<ipython-input-10-3aaf8ec2aad1>:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row\_indexer,col\_indexer] = value instead

df\_fake\_manual\_testing[“class”] = 0

<ipython-input-10-3aaf8ec2aad1>:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row\_indexer,col\_indexer] = value instead

df\_true\_manual\_testing[“class”] = 1

df\_fake\_manual\_testing.head(10)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **title** | **text** | **subject** | **date** | **class** |
| **23471** | Seven Iranians freed in the prisoner swap have... | 21st Century Wire says This week, the historic... | Middle-east | January 20, 2016 | 0 |
| **23472** | #Hashtag Hell & The Fake Left | By Dady Chery and Gilbert MercierAll writers ... | Middle-east | January 19, 2016 | 0 |
| **23473** | Astroturfing: Journalist Reveals Brainwashing ... | Vic Bishop Waking TimesOur reality is carefull... | Middle-east | January 19, 2016 | 0 |
| **23474** | The New American Century: An Era of Fraud | Paul Craig RobertsIn the last years of the 20t... | Middle-east | January 19, 2016 | 0 |
| **23475** | Hillary Clinton: ‘Israel First’ (and no peace ... | Robert Fantina CounterpunchAlthough the United... | Middle-east | January 18, 2016 | 0 |
| **23476** | McPain: John McCain Furious That Iran Treated ... | 21st Century Wire says As 21WIRE reported earl... | Middle-east | January 16, 2016 | 0 |
| **23477** | JUSTICE? Yahoo Settles E-mail Privacy Class-ac... | 21st Century Wire says It s a familiar theme. ... | Middle-east | January 16, 2016 | 0 |
| **23478** | Sunnistan: US and Allied ‘Safe Zone’ Plan to T... | Patrick Henningsen 21st Century WireRemember ... | Middle-east | January 15, 2016 | 0 |
| **23479** | How to Blow $700 Million: Al Jazeera America F... | 21st Century Wire says Al Jazeera America will... | Middle-east | January 14, 2016 | 0 |
| **23480** | 10 U.S. Navy Sailors Held by Iranian Military ... | 21st Century Wire says As 21WIRE predicted in ... | Middle-east | January 12, 2016 | 0 |
|  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **title** | **text** | **subject** | **date** | **class** |
| **21407** | Mata Pires, owner of embattled Brazil builder ... | SAO PAULO (Reuters) - Cesar Mata Pires, the ow... | worldnews | August 22, 2017 | 1 |
| **21408** | U.S., North Korea clash at U.N. forum over nuc... | GENEVA (Reuters) - North Korea and the United ... | worldnews | August 22, 2017 | 1 |
| **21409** | U.S., North Korea clash at U.N. arms forum on ... | GENEVA (Reuters) - North Korea and the United ... | worldnews | August 22, 2017 | 1 |
| **21410** | Headless torso could belong to submarine journ... | COPENHAGEN (Reuters) - Danish police said on T... | worldnews | August 22, 2017 | 1 |
| **21411** | North Korea shipments to Syria chemical arms a... | UNITED NATIONS (Reuters) - Two North Korean sh... | worldnews | August 21, 2017 | 1 |
| **21412** | 'Fully committed' NATO backs new U.S. approach... | BRUSSELS (Reuters) - NATO allies on Tuesday we... | worldnews | August 22, 2017 | 1 |
| **21413** | LexisNexis withdrew two products from Chinese ... | LONDON (Reuters) - LexisNexis, a provider of l... | worldnews | August 22, 2017 | 1 |
| **21414** | Minsk cultural hub becomes haven from authorities | MINSK (Reuters) - In the shadow of disused Sov... | worldnews | August 22, 2017 | 1 |
| **21415** | Vatican upbeat on possibility of Pope Francis ... | MOSCOW (Reuters) - Vatican Secretary of State ... | worldnews | August 22, 2017 | 1 |
| **21416** | Indonesia to buy $1.14 billion worth of Russia... | JAKARTA (Reuters) - Indonesia will buy 11 Sukh... | worldnews | August 22, 2017 | 1 |

df\_true\_manual\_testing.head(10)

df\_manual\_testing = pd.concat([df\_fake\_manual\_testing,df\_true\_manual\_testing], axis = 0)

df\_manual\_testing.to\_csv(“manual\_testing.csv”)

Merging True and Fake Dataframes:

df\_merge = pd.concat([df\_fake, df\_true], axis =0 )

df\_merge.head(10)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **title** | **text** | **subject** | **date** | **class** |
| **0** Donald Trump Sends Out Embarrassing New Year’... | Donald Trump just couldn t wish all Americans ... | News | December 31, 2017 | 0 |
| **1** Drunk Bragging Trump Staffer Started Russian ... | House Intelligence Committee Chairman Devin Nu... | News | December 31, 2017 | 0 |
| **2** Sheriff David Clarke Becomes An Internet Joke... | On Friday, it was revealed that former Milwauk... | News | December 30, 2017 | 0 |
| **3** Trump Is So Obsessed He Even Has Obama’s Name... | On Christmas day, Donald Trump announced that ... | News | December 29, 2017 | 0 |
| **4** Pope Francis Just Called Out Donald Trump Dur... | Pope Francis used his annual Christmas Day mes... | News | December 25, 2017 | 0 |
| **5** Racist Alabama Cops Brutalize Black Boy While... | The number of cases of cops brutalizing and ki... | News | December 25, 2017 | 0 |
| **6** Fresh Off The Golf Course, Trump Lashes Out A... | Donald Trump spent a good portion of his day a... | News | December 23, 2017 | 0 |
| **7** Trump Said Some INSANELY Racist Stuff Inside ... | In the wake of yet another court decision that... | News | December 23, 2017 | 0 |
| **8** Former CIA Director Slams Trump Over UN Bully... | Many people have raised the alarm regarding th... | News | December 22, 2017 | 0 |
| **9** WATCH: Brand-New Pro-Trump Ad Features So Muc... | Just when you might have thought we d get a br... | News | December 21, 2017 | 0 |

df\_merge.columns

index([‘title’,’text’,’subject’,’data’,’class’],dtype=’object’)

Removing columns which are not required

df = df\_merge.drop([“title”,” subject”,”date”], axis = 1)

df.isnull().sum()

text 0

class 0

dtype: int64

Random Shuffling the dataframe

df = df.sample(frac = 1)

df.head()

|  |  |  |
| --- | --- | --- |
|  | **text** | **class** |
| **7701** | Bobby Jindal abandoned his home state of Louis... | 0 |
| **16528** | Newsflash Hillary WAR is not aesthetically ple... | 0 |
| **9527** | WASHINGTON (Reuters) - The White House confirm... | 1 |
| **18516** | Texas Democrat Rep. Al Green announced on Wedn... | 0 |
| **993** | WASHINGTON (Reuters) - The U.S. State Departme... | 1 |

df.reset\_index(inplace = True)

df.drop([“index”], axis = 1, inplace = True)

df.columns

Index([‘text’, ‘class’], ‘object’)

df.head()

1. Bobby Jindal abandoned his home state of Louis... 0
2. Newsflash Hillary WAR is not aesthetically ple... 0
3. WASHINGTON (Reuters) - The White House confirm... 1
4. Texas Democrat Rep. Al Green announced on Wedn... 0
5. WASHINGTON (Reuters) - The U.S. State Departme... 1

def wordopt(text):

text = text.lower()

text = re.sub(‘\[.\*?\]’, &#39;&#39;, text)

text = re.sub(&quot;\\W&quot;,&quot; &quot;,text)

text = re.sub(&#39;https?://\S+|www\.\S+&#39;, &#39;&#39;, text)

text = re.sub(&#39;&lt;.\*?&gt;+&#39;, &#39;&#39;, text)

text = re.sub(&#39;[%s]&#39; % re.escape(string.punctuation), &#39;&#39;, text)

text = re.sub(&#39;\n&#39;, &#39;&#39;, text)

text = re.sub(&#39;\w\*\d\w\*&#39;, &#39;&#39;, text)

return text

df[“text”] = df[“text”].apply(wordopt)

Defining dependent and independent variables

x = df[“text”]

y = df[“class”]

**CONCLUTION:**

In conclusion, this research underscores the importance of preprocessing in enhancing data quality and the essential role of meticulously curated datasets in the advancement of fake news detection. The methods and insights presented here contribute to the ongoing efforts to combat misinformation in the digital age.