**GROUP 5 - Fake News Detection Using NLP**

**Phase 3: Collect datasets , Preprocessing Methods, Coding for preprocessing methods.**

1. **Datasets for COVID19 Fake News using NLP**:

Label tweet

1 Our daily update is published. States reported 734k tests 39k new cases and 532 deaths. Current hospitalizations fell below 30k for the first time since June 22. https://t.co/wzSYMe0Sht

2 Alfalfa is the only cure for COVID-19.

3 President Trump Asked What He Would Do If He Were To Catch The Coronavirus https://t.co/3MEWhusRZI #donaldtrump #coronavirus

4 States reported 630 deaths. We are still seeing a solid national decline. Death reporting lags approximately 28 days from symptom onset according to CDC models that consider lags in symptoms time in hospital and the death reporting process. https://t.co/LBmcot3h9a

5 This is the sixth time a global health emergency has been declared under the International Health Regulations but it is easily the most severe-@DrTedros https://t.co/JvKC0PTett

6 Low #vitaminD was an independent predictor of worse prognosis in patients with COVID-19. https://t.co/CGD6Kphn31 https://t.co/chtni8K4Jd

7 A common question: why are the cumulative outcome numbers smaller than the current outcome numbers? A: Most states report current but a few states report cumulative. They are apples and oranges and we don't feel comfortable filling in state cumulative boxes with current #s.

8 The government should consider bringing in any new national lockdown rules over Christmas rather than now says an Oxford University professor https://t.co/pdOls6cqoN

9 Our daily update is published. Weâ€™ve now tracked more than 2.9 million tests up 119k from yesterday. That's the smallest reported increase since April 1. Note that we can only track tests that a state reports. And not all states report all tests. See: https://t.co/PZrmH4bl5Y https://t.co/2588xW5yNm

10 Breakdown of testing: 4 air crew 97 hotel &amp; health staff in the facility 71 former hotel guests &amp; 2 exempted individuals who returned negative tests. And there are over 200 current guests in the Novotel who were swabbed on Tuesday &amp; Wednesday. Their results were negative.â£

11 "Two interesting correlations:

1) Children tend to weather COVID-19 pretty well; they also get a ton of Vitamin D.

2) Black people are getting slammed by COVID-19; black people also have much higher instances of Vitamin D deficiency (76% vs 40% in the general population)."

12 A photo shows a 19-year-old vaccine for canine coronavirus that could be used to prevent the new coronavirus causing COVID-19.

13 ðŸ‡°ðŸ‡¼ Assistant Undersecretary for Public Health Affairs Dr Buthayna Almodaf highlighted the importance of risk communication &amp; public education in fighting #COVID19. The country increased testing capacity enabling 400000 people to be tested.

14 An audio file by an alleged worker at a health institution in Rio de Janeiro. She says that healthcare workers on public institutions in Rio are forced to state whether a patient has COVID-19 or not even before he sees a doctor. This was allegedly being done to artificially inflate the number of cases.

15 Says the Coronavirus Aid, Relief, and Economic Security Act gives members of Congress a pay increase.

16 Kids reach â€˜f\*\*k this shitâ€™ stage of lockdown https://t.co/swClEhi2Iq

17 Households should have "required" medical kits with certain items and equipment to treat the different stages of COVID-19.

18 An image of a man carrying his old mother on his back shows migrants travelling to their homes in India amidst lockdown.

19 While #socialdistancing for #COVID19 take the time to check in on friends and family to see how they are doing. Set up a daily phone or video call to touch base and share the best and most challenging parts of the day. #TogetherApart https://t.co/KyBmehvGcY https://t.co/rN1RmYS9i0

20 3/10 About 8% of population "may be infected and have antibodies". The "vast majority" of people are still vulnerable to the disease says Sir Patrick Vallance. https://t.co/SWcqc3XLyJ https://t.co/drmAv7kmBA

21 "Korona virus, very new deadly form of virus, china is suffering, may come to

India immediately, avoid any form of cold drinks, ice creams, koolfee,

etc, any type of preserved foods, milkshake, rough ice, ice colas,

milk sweets older then 48 hours, for atleast 90 days from today."

22 A 2009 tweet from Donald Trump said that he would never let thousands of Americans die from a pandemic while in office.

23 ICUs are full forcing #COVID patients to wait â€œhours and hoursto admit to the emergency room. https://t.co/go6Sx5yuRx

24 President Donald Trump tweeted, "Coronavirus vaccine from Israel! Stock market up! Economy even stronger now! Besides US, Israel best at inventions: Judaism, olives. Wonderful vaccine, will stop Corona DEAD around world. Can't make vaccines when being bombed, not have been possible without peace I created in Israel!"

25 New York continues to have the highest positive tests per capita an indication of both the intensity of testing there and the severity of the outbreak. Hereâ€™s the top 10: New York Washington New Jersey Louisiana DC Michigan Illinois Vermont Colorado Rhode Island

26 The government's chief scientific adviser warned that the UK could soon see 49000 new cases of #coronavirus every day unless action is taken. Sky's @EdConwaySky takes a look at the data and suggests the real number could be lower. Latest: https://t.co/kORpcHXQIJ https://t.co/TdNeKOAlCq

27 Brazil's health minister says cases with no confirmation of COVID-19 won't be considered on death toll.

28 "Govt has added #Corona disease in all existing mediclaim insurances as a special case

#COVID2019India https://t.co/39vpW7tBqq"

29 She has been in managed Isolation at the Distinction Hotel in Hamilton and tested positive around day 3 of her stay. â€‹ We have 13 new confirmed cases in Auckland. They are all linked to the 4 people who we reported as cases yesterday and the day before.

30 The Gates Foundation "has the patent for this coronavirus."

31 Nashville Man Thinks World Is Upside-Down but Respects that World Doesn't Care What He Thinks https://t.co/xAw4YLwLsl #coronavirus

32 14 China Muslims hidden at Bihari mosque has been taken to corona virus test by Bihari police. Erode police has caught Thailand Muslim mullahs infected with corona virus. Today Salem Police has caught 11 Indonesian Muslim mullahs at Salem mosque. This video shows that they are applying and putting saliva on spoons, plates and utensils and also they are in the intention of spreading corona virus disease. Nobody knows what's happening in the Nation

33 There are also clear \*demand\* constraints as testing criteria remain quite strict. We are still getting reports from all over of people who have been able unable to get tested despite having a good reason to do so.

34 Coronavirus found in gutkha or chewing tobacco samples sold in Indian state of Maharashtra as raw material sourced from China.

35 Clinical senior lecturer at King's College London Dr Claire Steves says 'two symptoms of #COVID19 that are not on the core list are headaches and fatigue' and discusses what symptoms warrant a test for #coronavirus ðŸ‘‡ Latest: https://t.co/gpqYW8xYoB https://t.co/nmvIiVwZ3P

36 @FriedrichHayek @NateSilver538 Not meaningless but also before this epidemic pneumonia deaths were considered a very unstable number... so keep that in mind. -@alexismadrigal

37 @LimaEleanor Kia ora Elanor there are currently 21 active cases in New Zealand they were all caught at the border and we have no evidence of community transmission.

38 Update from the @minhealthnz â€‹Today there are 3 new cases of COVID-19 to report in managed isolation and quarantine facilities in NZ. There continue to be no new cases in the community.â€‹ Our total number of active cases is 21 - all of which remain in quarantine facilities.

39 The new coronavirus does not settle in the air but is grounded so it is not transmitted by air.

40 If you have signs of a heart attack or stroke while staying at home during #COVID19 call 9-1-1 right away. Emergency responders and emergency departments have plans in place to help protect you from COVID-19. Learn more: https://t.co/4i0e7Cw3m1. https://t.co/ZqLIrVCHXI

41 The Access to #COVID19 Tools (ACT) Accelerator brings together governments health organizations scientists businesses civil society &amp; philanthropists to speed up an end to the pandemic. ðŸ‘‰https://t.co/fXHNTOsLVR https://t.co/kTXtFlm2jU

42 Says the CDC now says that the coronavirus can survive on surfaces for up to 17 days.

43 The Chinese government announced that "garlic is a preventative food for the the novel coronavirus."

44 Flood Hit Areas Pretend To Have Coronavirus In Bid To Get Government's Attention https://t.co/NKpkH0cY0Z

45 @Dianes\_Bucket @arkless\_kevin @ReeeeSpectWalk @ftwrharry @wolfiecindy 'Spanish Flu' was propaganda to cover up the millions of soldiers who died from military vaccines, WWI war trauma, radiation effects and hysteria (lots of Aspirin overdose by people who thought it would protect them).

46 "ðŸ¤ As business continue to adapt their services in the wake of covid-19, the SWA Hand Sanitiser Portal is available for sourcing #handsanitiser and ethanol for producing sanitiser.

ðŸ–Šï¸ Complete your details and find suppliers here: https://t.co/b7uLQbfQLz https://t.co/5qJQ5h8w6l"

47 FREE CE: Watch the new #CDCGrandRounds on Acute Flaccid Myelitis (AFM). Listen to our new #BeyondtheData podcast as CDCâ€™s Dr. John Iskander and Dr. Janell Routh discuss AFM its symptoms and treatment options. https://t.co/uD0xXEe4PR https://t.co/VtcvnAanq7

48 "Hydroxychloroquine. The virus cure.

How fast does this pandemic end if you could start feeling better in as soon as 5 hours?"

49 SUBHAN ALLAH: AFTER CORONA VIRUS CHINA GOVT LIFTED BAN ON HOLY QURAN & ALLOWED CHINESE MUSLIMS TO READ THEIR SACRED BOOK! "SO WHICH OF THE FAVORS OF YOUR LORD WOULD YOU DENY?"

50 Our daily update is published. States reported 671k tests 40k cases and 1202 COVID-19 deaths. Note: we've added 7-day averages to the nationwide overview. https://t.co/BQP32jcUi1

**2.Preprocessing Methods:**

Preprocessing text data for natural language processing (NLP) tasks, such as classifying COVID-19 fake news, involves a series of steps to clean and transform the text into a format suitable for machine learning models. Here's a comprehensive list of preprocessing methods specifically tailored for COVID-19 fake news datasets:

1. Lowercasing:

Convert all text to lowercase to ensure uniformity in the dataset.

2. Removing URLs and Special Characters:

Remove URLs, hashtags, mentions, and other special characters from the text as they do not contribute to the meaning.

3. Tokenization:

Tokenize the text to break it down into individual words or tokens. This step is essential for further text processing.

4. Removing Stopwords:

Remove common stopwords (e.g., 'and', 'the', 'is') as they do not carry significant meaning.

5. Removing Non-Alphabetic Characters:Remove numbers and other non-alphabetic characters that might not be relevant for text classification.

6. Lemmatization or Stemming:

Reduce words to their base or root form. Lemmatization is preferred as it ensures resulting words are valid, but stemming can also be used for faster processing.

7. Handling Emojis and Emoticons:

Decide whether to keep, remove, or replace emojis and emoticons based on the context of your analysis.

8. Handling Short Words:

Decide whether to keep or remove very short words (e.g., words with one or two characters) based on your analysis goals.

9. Handling Negations:

Consider adding a prefix like 'not\_' to words appearing after negations (e.g., 'not good'). This step is essential for sentiment analysis tasks.

10. Handling Spelling Errors:

sqlCorrect common spelling mistakes using libraries like TextBlob or spell checkers.

11. Vectorization:

vbnet

Convert the preprocessed text into numerical vectors using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (Word2Vec, GloVe). This step is necessary for most machine learning models to process

**Coding:**

import re

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer

# import nltk

# nltk.download('punkt')

# nltk.download('stopwords')

# nltk.download('wordnet')

# Function to preprocess text

def preprocess\_text(text):

# Convert to lowercase

text = text.lower()

# Remove URLs, special characters, and numbers

text = re.sub(r'http\S+|www\S+|https\S+|[^a-zA-Z]', ' ', text)

# Tokenization

tokens = word\_tokenize(text)

# Remove stopwords

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

filtered\_tokens = [word for word in tokens if word not in stop\_words]

# Lemmatization

lemmatizer = WordNetLemmatizer()

lemmatized\_tokens = [lemmatizer.lemmatize(word) for word in filtered\_tokens]

# Join the tokens back into text

preprocessed\_text = ' '.join(lemmatized\_tokens)

return preprocessed\_text

# Apply preprocessing to the 'tweet' column

df['cleaned\_tweet'] = df['tweet'].apply(preprocess\_text)

# Now df['cleaned\_tweet'] contains the preprocessed textimport pandas as pd

data = {'tweet': ['Hello, World! Visit https://example.com for more info!',

'This is a sample tweet with some numbers 123 and special characters #$%@']}

df = pd.DataFrame(data)

O/P:

tweet cleaned\_tweet

0 Hello, World! hello world

1 This is a tweet with some numbers 123 tweet number special characters.

Coding for COVID19 Fake News Dataset NLP& Preprocessing:

import pandas as pd

# Load the dataset into a DataFrame

df=pd.read\_csv(r'C:\Users\DELL\Documents\datasets\csvfile\Constraint\_Test.csv')

# Explore the first few rows of the dataset

print(df.head())

# Assuming 'tweet' is the column containing the text and 'label' is the column containing labels

# Removing unnecessary columns

df = df[['tweet', 'label']]

# Checking for and handling missing values

df.dropna(inplace=True)

# Converting labels to binary (1 for fake news, 0 for real news)

df['label'] = df['label'].apply(lambda x: 1 if x > 0 else 0)

# Text cleaning (you might need more advanced text preprocessing steps)

import re

def clean\_text(text):

text = re.sub(r'http\S+', '', text) # Remove URLs

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

text = text.lower() # Convert to lowercase

return text

df['cleaned\_tweet'] = df['tweet'].apply(clean\_text)

from sklearn.feature\_extraction.text import TfidfVectorizer

# Tokenization and vectorization

tfidf\_vectorizer = TfidfVectorizer(stop\_words='english', max\_features=5000)

X = tfidf\_vectorizer.fit\_transform(df['cleaned\_tweet'])

y = df['label']

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

from sklearn.naive\_bayes import MultinomialNB

from sklearn.metrics import accuracy\_score, classification\_report

# Split the data into training and testing sets

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

# Train a Naive Bayes classifier

clf = MultinomialNB()

clf.fit(X\_train, y\_train)

# Predictions on the test set

predictions = clf.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, predictions)

print(f'Accuracy: {accuracy:.2f}')

print(classification\_report(y\_test, predictions))

**Output:**

label tweet

0 1 Our daily update is published. States reported...

1 2 Alfalfa is the only cure for COVID-19.

2 3 President Trump Asked What He Would Do If He W...

3 4 States reported 630 deaths. We are still seein...

4 5 This is the sixth time a global health emergen...

Accuracy: 1.00

precision recall f1-score support

1 1.00 1.00 1.00 428

accuracy 1.00 428

macro avg 1.00 1.00 1.00 428

weighted avg 1.00 1.00 1.00 428