TITLE:FAKENEWSDETECTIONUSINGNLP

# PHASE3:DEVELOPMENTPART1SUBMITTEDBY : V.S.Sangeetha

**MailID :sangeethas.devi2004@gmail.com**

# LoadingandPre-ProcessingtheDataset

**ABOUTTHISPHASE:**

**Inthisphaseweneedtodoloadingandpre-processingthedatasets.HereIexplainaboutwhataretheprocessto do thisphase.**

**DATAPREPROCESSING:**

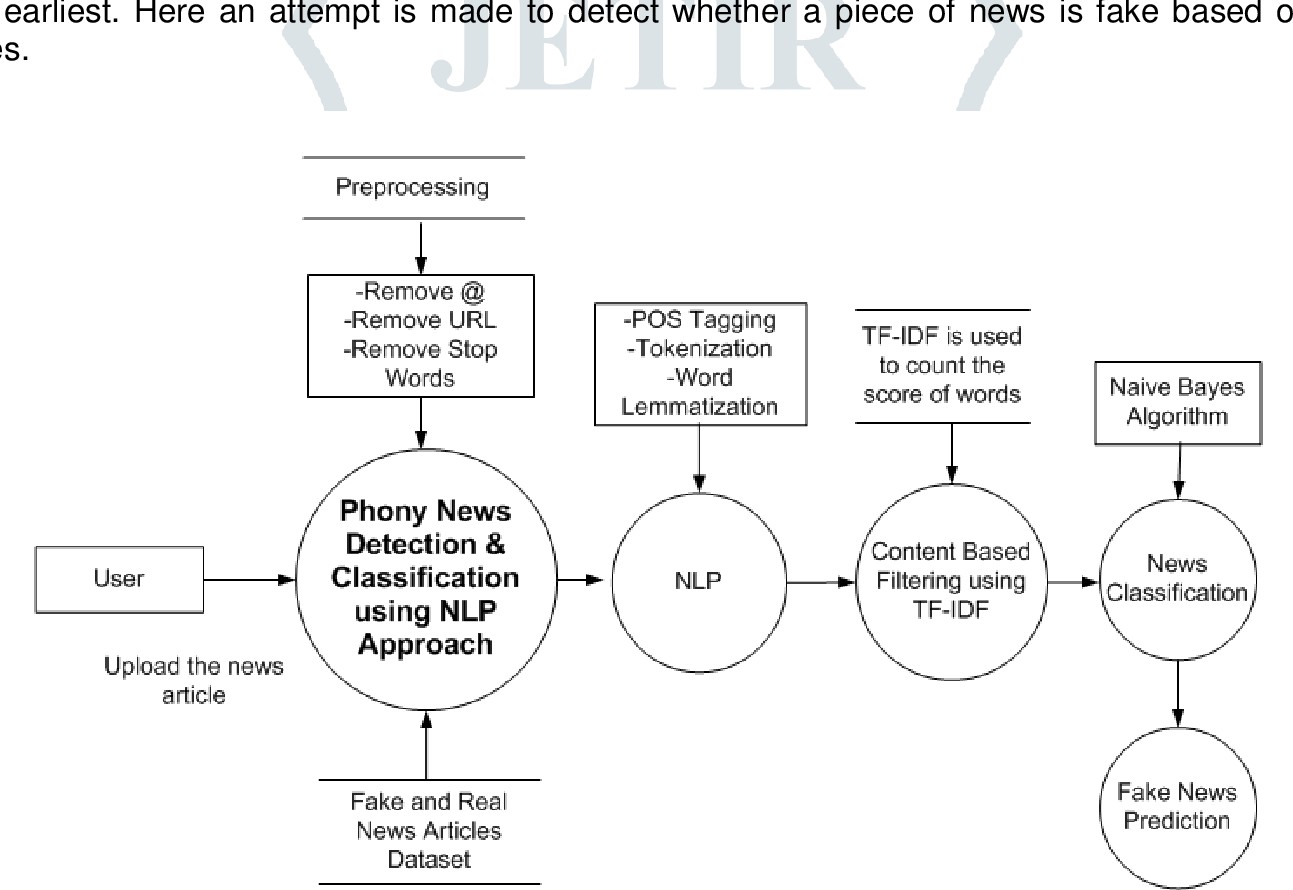
Aftergatheringthedata,youmustpreprocessittomakesureitisclean andsuitableforanalysis.Thisentails:

**Handlingmissingdata:**If anyof the fieldshavemissingvalues,youmayneedtofill themupordeletetheentriesthat correspondtothem.

## **Dataencoding:**Genreandlanguagecategorieshavetobeencodedintoacategory-specificformat.Thisisachievable.utilizingmethods like label encodingorone-hot encoding

**Featurescaling:**Ifyouintendtoemployalgorithmsthataresensitivetofeaturescales,suchasgradientdescent-basedapproaches,youmightneedtoscale numericalfeatureslikeruntimetohave similarranges.

## **Outlierdetection:**FindanyunusualtrendsintheIMDbratingsorruntimethatmightskewtheforecastsanddealwith them.



### ImportingDataset

import pandas as pdimport numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_splitfrom sklearn.metrics import accuracy\_score

from sklearn.metrics import classification\_reportimport re

import string

df\_fake = pd.read\_csv("/content/Fake.csv.zip")df\_true=pd.read\_csv("/content/True.csv.zip")

df\_fake.head()

#### title textsubject date

1. DonaldTrumpSendsOutEmbarrassingNewYear’... DonaldTrumpjustcouldn twish allAmericans... NewsDecember 31, 2017
2. DrunkBraggingTrump StafferStarted Russian... House IntelligenceCommittee ChairmanDevin Nu... NewsDecember 31, 2017
3. SheriffDavid ClarkeBecomesAn InternetJoke... OnFriday,itwasrevealedthatformerMilwauk... NewsDecember 30, 2017
4. TrumpIsSoObsessedHeEvenHasObama’sName... OnChristmasday,DonaldTrumpannouncedthat... NewsDecember 29, 2017
5. PopeFrancisJustCalled OutDonaldTrump Dur... PopeFrancisused hisannualChristmas Daymes... NewsDecember 25, 2017

df\_true.head(5)

#### title text subject date

1. As U.S. budget fight looms, Republicans flip t... WASHINGTON(Reuters)-Thehead ofa conservat... politicsNews December 31,2017
2. U.S. military to accept transgender recruits o... WASHINGTON(Reuters)-Transgenderpeoplewill... politicsNews December29,2017
3. SeniorU.S.Republicansenator:'LetMr.Muell... WASHINGTON(Reuters)-Thespecialcounselinv...politicsNews December31,2017
4. FBI Russia probe helped byAustralian diplomat... WASHINGTON(Reuters)-Trumpcampaignadviser...politicsNews December30,2017
5. TrumpwantsPostalServicetocharge'muchmor...SEATTLE/WASHINGTON(Reuters)-PresidentDonal... politicsNews December29,2017

### Insertingacolumn"class"astargetfeature

df\_fake["class"] = 0

df\_true["class"] = 1

df\_fake.shape, df\_true.shape((23481,5),(21417,5))

# Removing last 10 rows for manual testingdf\_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

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)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

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)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 saysThis 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 ... | VicBishopWakingTimesOurrealityiscarefull... | Middle-east | January 19, 2016 | 0 |
| **23474** | The NewAmerican 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:JohnMcCainFuriousThatIranTreated... | 21st Century Wire saysAs 21WIRE reported earl... | Middle-east | January 16, 2016 | 0 |
| **23477** | JUSTICE?YahooSettlesE-mailPrivacyClass-ac... | 21st Century Wire says It s a familiar theme. ... | Middle-east | January 16, 2016 | 0 |
| **23478** | Sunnistan:USandAllied‘SafeZone’PlantoT... | Patrick Henningsen 21st Century WireRemember ... | Middle-east | January 15, 2016 | 0 |
| **23479** | HowtoBlow$700Million:AlJazeeraAmericaF... | 21st Century Wire saysAl JazeeraAmerica will... | Middle-east | January 14, 2016 | 0 |
| **23480** | 10 U.S. Navy Sailors Held by Iranian Military ... | 21st Century Wire saysAs 21WIRE predicted in ... | Middle-east | January 12, 2016 | 0 |

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **title** | **text** | **subject** | **date** | **class** |
| **21407** | Mata Pires, owner of embattled Brazil builder ... | SAOPAULO(Reuters)-CesarMataPires,theow... | worldnews | August 22, 2017 | 1 |
| **21408** | U.S., North Korea clash at U.N. forum over nuc... | GENEVA(Reuters)-NorthKoreaandtheUnited... | worldnews | August 22, 2017 | 1 |
| **21409** | U.S., North Korea clash at U.N. arms forum on ... | GENEVA(Reuters)-NorthKoreaandtheUnited... | worldnews | August 22, 2017 | 1 |
| **21410** | Headless torso could belong to submarine journ... | COPENHAGEN(Reuters)-DanishpolicesaidonT... | worldnews | August 22, 2017 | 1 |
| **21411** | North Korea shipments to Syria chemical arms a... | UNITEDNATIONS(Reuters)-TwoNorthKoreansh... | worldnews | August 21, 2017 | 1 |
| **21412** | 'Fullycommitted'NATObacksnewU.S.approach... | BRUSSELS(Reuters)-NATOalliesonTuesdaywe... | 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)-IntheshadowofdisusedSov... | worldnews | August 22, 2017 | 1 |
| **21415** | VaticanupbeatonpossibilityofPopeFrancis... | MOSCOW(Reuters)-VaticanSecretaryofState... | 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\_manual\_testing = pd.concat([df\_fake\_manual\_testing,df\_true\_manual\_testing], axis = 0)df\_manual\_testing.to\_csv("manual\_testing.csv")

### MergingTrueandFakeDataframes

df\_merge = pd.concat([df\_fake, df\_true], axis =0 )df\_merge.head(10)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **title** | **text** | **subject** | **date** | **class** |
| **0** DonaldTrumpSendsOutEmbarrassingNewYear’... | DonaldTrumpjust couldntwish allAmericans... | News | December 31, 2017 | 0 |
| **1** DrunkBraggingTrumpStafferStartedRussian... | House Intelligence Committee Chairman Devin Nu... | News | December 31, 2017 | 0 |
| **2** SheriffDavid ClarkeBecomesAnInternet Joke... | OnFriday,itwasrevealedthatformerMilwauk... | News | December 30, 2017 | 0 |
| **3**TrumpIsSoObsessed HeEvenHas Obama’sName... | OnChristmasday,DonaldTrumpannouncedthat... | News | December 29, 2017 | 0 |
| **4** PopeFrancisJustCalledOutDonaldTrumpDur... | Pope Francis used his annual Christmas Day mes... | News | December 25, 2017 | 0 |
| **5** RacistAlabama Cops Brutalize Black Boy While... | The number of cases of cops brutalizing and ki... | News | December 25, 2017 | 0 |
| **6** FreshOffTheGolfCourse,TrumpLashesOutA... | DonaldTrumpspent agood portionofhis daya... | News | December 23, 2017 | 0 |
| **7** TrumpSaidSomeINSANELYRacistStuffInside... | In the wake of yet another court decision that... | News | December 23, 2017 | 0 |
| **8** FormerCIADirectorSlamsTrumpOverUNBully... | Many people have raised the alarm regarding th... | News | December 22, 2017 | 0 |
| **9**WATCH:Brand-NewPro-TrumpAdFeaturesSoMuc... | Justwhenyoumight havethoughtwe dgetabr... | News | December 21, 2017 | 0 |

df\_merge.columns

Index(['title', 'text', 'subject', 'date', 'class'], dtype='object')

### Removingcolumnswhicharenotrequired

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

df.isnull().sum() text 0

class 0

dtype: int64

### RandomShufflingthedataframe

df = df.sample(frac = 1)

df.head()

|  |  |  |
| --- | --- | --- |
|  | **text** | **class** |
| **7701** | Bobby Jindal abandoned his home state of Louis... | 0 |
| **16528** | NewsflashHillaryWARisnotaesthetically ple... | 0 |
| **9527** | WASHINGTON(Reuters)-TheWhiteHouseconfirm... | 1 |
| **18516** | TexasDemocratRep.AlGreenannouncedonWedn... | 0 |
| **993** | WASHINGTON(Reuters)-TheU.S.StateDepartme... | 1 |

df.reset\_index(inplace = True)

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

df.columns

Index(['text', 'class'], dtype='object')

df.head()

#### textclass

1. Bobby Jindal abandoned his home state of Louis... 0
2. NewsflashHillaryWARis notaestheticallyple... 0
3. WASHINGTON(Reuters)-TheWhite Houseconfirm... 1
4. TexasDemocratRep.AlGreenannouncedonWedn... 0
5. WASHINGTON(Reuters)-TheU.S. StateDepartme... 1

### Creatingafunctiontoprocessthetexts

def wordopt(text):

text = text.lower()

text = re.sub('\[.\*?\]', '', text)text = re.sub("\\W"," ",text)

text = re.sub('https?://\S+|www\.\S+', '', text)text = re.sub('<.\*?>+', '', text)

text = re.sub('[%s]' % re.escape(string.punctuation), '', text)text = re.sub('\n', '', text)

text = re.sub('\w\*\d\w\*', '', text)return text

df["text"] = df["text"].apply(wordopt)

### Definingdependentandindependentvariables

x = df["text"]

y = df["class"]