AI_Phase3

October 14, 2023

Build an NLP model to differentiate real news from fake news

Phase 3: Development Part 1 .

In this part you will begin building your project by loading and preprocessing the dataset. Begin building the fake news detection model by loading and preprocessing the dataset. Load the fake news dataset and preprocess the textual data.

`Dataset Link: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset`

Importing required libraries

[2]:

```
import warnings
warnings.filterwarnings('ignore')
```

[3]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import nltk
import re
import string

from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report

import keras
from keras.preprocessing import text,sequence
from keras.models import Sequential
from keras.layers import Dense,Embedding,LSTM,Dropout

import os
for dirname, _, filenames in os.walk('dataset/'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
dataset/Fake.csv
dataset/archive(1).zip
dataset/True.csv
```

Loading Data

[4]:
```
real_data = pd.read_csv('dataset/True.csv')
fake_data = pd.read_csv('dataset/Fake.csv')
```

[5]:
```
real_data.head()
```

[5]:
```
                                               title  \
0  As U.S. budget fight looms, Republicans flip t…
1  U.S. military to accept transgender recruits o…
2  Senior U.S. Republican senator: 'Let Mr. Muell…
3  FBI Russia probe helped by Australian diplomat…
4  Trump wants Postal Service to charge 'much mor…

                                                text       subject  \
0  WASHINGTON (Reuters) - The head of a conservat…  politicsNews
1  WASHINGTON (Reuters) - Transgender people will…  politicsNews
2  WASHINGTON (Reuters) - The special counsel inv…  politicsNews
3  WASHINGTON (Reuters) - Trump campaign adviser …  politicsNews
4  SEATTLE/WASHINGTON (Reuters) - President Donal…  politicsNews

                 date
0  December 31, 2017
1  December 29, 2017
2  December 31, 2017
3  December 30, 2017
4  December 29, 2017
```

[6]:
```
fake_data.head()
```

[6]:
```
                                               title  \
0   Donald Trump Sends Out Embarrassing New Year’…
1   Drunk Bragging Trump Staffer Started Russian …
2   Sheriff David Clarke Becomes An Internet Joke…
3   Trump Is So Obsessed He Even Has Obama’s Name…
4   Pope Francis Just Called Out Donald Trump Dur…

                                                text subject  \
0  Donald Trump just couldn t wish all Americans …    News
1  House Intelligence Committee Chairman Devin Nu…    News
2  On Friday, it was revealed that former Milwauk…    News
3  On Christmas day, Donald Trump announced that …    News
4  Pope Francis used his annual Christmas Day mes…    News

                date
0  December 31, 2017
1  December 31, 2017
2  December 30, 2017
```

```
3  December 29, 2017
4  December 25, 2017
```

[7]:

```python
#add column
real_data['target'] = 1
fake_data['target'] = 0
```

[8]:

```python
real_data.tail()
```

[8]:

```
                                                   title  \
21412  'Fully committed' NATO backs new U.S. approach…
21413  LexisNexis withdrew two products from Chinese …
21414  Minsk cultural hub becomes haven from authorities
21415  Vatican upbeat on possibility of Pope Francis …
21416  Indonesia to buy $1.14 billion worth of Russia…

                                                    text    subject  \
21412  BRUSSELS (Reuters) - NATO allies on Tuesday we…  worldnews
21413  LONDON (Reuters) - LexisNexis, a provider of l…  worldnews
21414  MINSK (Reuters) - In the shadow of disused Sov…  worldnews
21415  MOSCOW (Reuters) - Vatican Secretary of State …  worldnews
21416  JAKARTA (Reuters) - Indonesia will buy 11 Sukh…  worldnews

                   date  target
21412  August 22, 2017        1
21413  August 22, 2017        1
21414  August 22, 2017        1
21415  August 22, 2017        1
21416  August 22, 2017        1
```

[9]:

```python
#Merging the 2 datasets
data = pd.concat([real_data, fake_data], ignore_index=True, sort=False)
data.head()
```

[9]:

```
                                               title  \
0  As U.S. budget fight looms, Republicans flip t…
1  U.S. military to accept transgender recruits o…
2  Senior U.S. Republican senator: 'Let Mr. Muell…
3  FBI Russia probe helped by Australian diplomat…
4  Trump wants Postal Service to charge 'much mor…

                                                text       subject  \
0  WASHINGTON (Reuters) - The head of a conservat…  politicsNews
1  WASHINGTON (Reuters) - Transgender people will…  politicsNews
2  WASHINGTON (Reuters) - The special counsel inv…  politicsNews
3  WASHINGTON (Reuters) - Trump campaign adviser …  politicsNews
4  SEATTLE/WASHINGTON (Reuters) - President Donal…  politicsNews
```

```
                 date  target
0  December 31, 2017       1
1  December 29, 2017       1
2  December 31, 2017       1
3  December 30, 2017       1
4  December 29, 2017       1
```

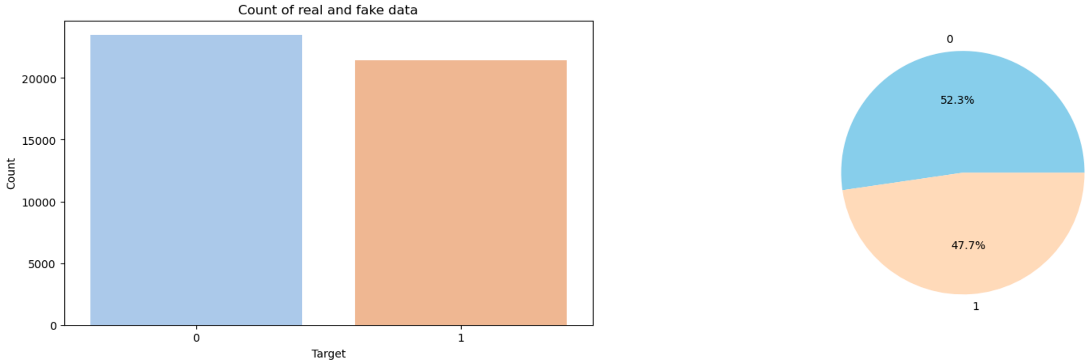
```
[10]: data.isnull().sum()
```

```
[10]: title      0
      text       0
      subject    0
      date       0
      target     0
      dtype: int64
```

1.Count of Fake and Real Data

```
[11]: print(data["target"].value_counts())
      fig, ax = plt.subplots(1,2, figsize=(19, 5))
      g1 = sns.countplot(data.target,ax=ax[0],palette="pastel");
      g1.set_title("Count of real and fake data")
      g1.set_ylabel("Count")
      g1.set_xlabel("Target")
      g2 = plt.pie(data["target"].value_counts().values,explode=[0,0],labels=data.
        ↪target.value_counts().index, autopct='%1.
        ↪1f%%',colors=['SkyBlue','PeachPuff'])
      fig.show()
```

```
0    23481
1    21417
Name: target, dtype: int64
```



2.Distribution of The Subject According to Real and Fake Data

[12]:

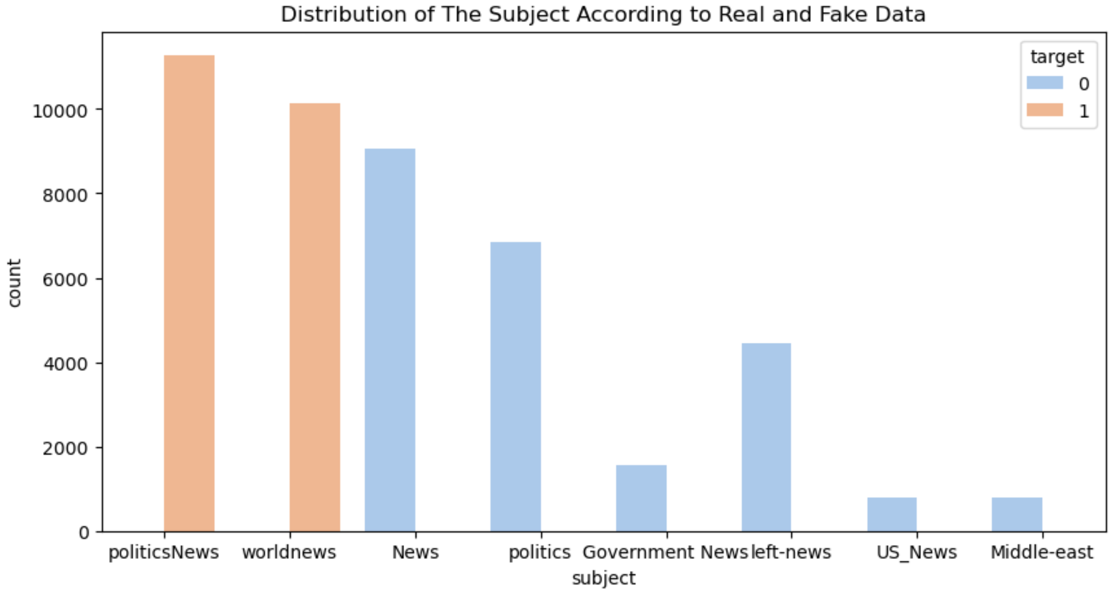
```
print(data.subject.value_counts())
plt.figure(figsize=(10, 5))

ax = sns.countplot(x="subject",  hue='target', data=data, palette="pastel")
plt.title("Distribution of The Subject According to Real and Fake Data")
```

```
politicsNews       11272
worldnews          10145
News                9050
politics            6841
left-news           4459
Government News     1570
US_News              783
Middle-east          778
Name: subject, dtype: int64
```

[12]:

```
Text(0.5, 1.0, 'Distribution of The Subject According to Real and Fake Data')
```



Preprocessing the textual data

[13]:

```
data['text']= data['subject'] + " " + data['title'] + " " + data['text']
del data['title']
del data['subject']
del data['date']
data.head()
```

[13]:

```
                                                text  target
0  politicsNews As U.S. budget fight looms, Repub…       1
```

```
1  politicsNews U.S. military to accept transgend…       1
2  politicsNews Senior U.S. Republican senator: '…       1
3  politicsNews FBI Russia probe helped by Austra…       1
4  politicsNews Trump wants Postal Service to cha…       1
```

[14]:
```python
first_text = data.text[10]
first_text
```

[14]:
```
'politicsNews Jones certified U.S. Senate winner despite Moore challenge
(Reuters) - Alabama officials on Thursday certified Democrat Doug Jones the
winner of the state's U.S. Senate race, after a state judge denied a challenge
by Republican Roy Moore, whose campaign was derailed by accusations of sexual
misconduct with teenage girls. Jones won the vacant seat by about 22,000 votes,
or 1.6 percentage points, election officials said. That made him the first
Democrat in a quarter of a century to win a Senate seat in Alabama.  The seat
was previously held by Republican Jeff Sessions, who was tapped by U.S.
President Donald Trump as attorney general. A state canvassing board composed of
Alabama Secretary of State John Merrill, Governor Kay Ivey and Attorney General
Steve Marshall certified the election results. Seating Jones will narrow the
Republican majority in the Senate to 51 of 100 seats. In a statement, Jones
called his victory "a new chapter" and pledged to work with both parties. Moore
declined to concede defeat even after Trump urged him to do so. He stood by
claims of a fraudulent election in a statement released after the certification
and said he had no regrets, media outlets reported. An Alabama judge denied
Moore's request to block certification of the results of the Dec. 12 election in
a decision shortly before the canvassing board met. Moore's challenge alleged
there had been potential voter fraud that denied him a chance of victory. His
filing on Wednesday in the Montgomery Circuit Court sought to halt the meeting
scheduled to ratify Jones' win on Thursday. Moore could ask for a recount, in
addition to possible other court challenges, Merrill said in an interview with
Fox News Channel. He would have to complete paperwork "within a timed period"
and show he has the money for a challenge, Merrill said. "We've not been
notified yet of their intention to do that," Merrill said. Regarding the claim
of voter fraud, Merrill told CNN that more than 100 cases had been reported.
"We've adjudicated more than 60 of those. We will continue to do that," he said.
Republican lawmakers in Washington had distanced themselves from Moore and
called for him to drop out of the race after several women accused him of sexual
assault or misconduct dating back to when they were teenagers and he was in his
early 30s.  Moore has denied wrongdoing and Reuters has not been able to
independently verify the allegations. '
```

First, let's remove HTML content.

[15]:
```python
from bs4 import BeautifulSoup

soup = BeautifulSoup(first_text, "html.parser")
first_text = soup.get_text()
first_text
```

```
[15]: 'politicsNews Jones certified U.S. Senate winner despite Moore challenge
      (Reuters) - Alabama officials on Thursday certified Democrat Doug Jones the
      winner of the state’s U.S. Senate race, after a state judge denied a challenge
      by Republican Roy Moore, whose campaign was derailed by accusations of sexual
      misconduct with teenage girls. Jones won the vacant seat by about 22,000 votes,
      or 1.6 percentage points, election officials said. That made him the first
      Democrat in a quarter of a century to win a Senate seat in Alabama.  The seat
      was previously held by Republican Jeff Sessions, who was tapped by U.S.
      President Donald Trump as attorney general. A state canvassing board composed of
      Alabama Secretary of State John Merrill, Governor Kay Ivey and Attorney General
      Steve Marshall certified the election results. Seating Jones will narrow the
      Republican majority in the Senate to 51 of 100 seats. In a statement, Jones
      called his victory "a new chapter" and pledged to work with both parties. Moore
      declined to concede defeat even after Trump urged him to do so. He stood by
      claims of a fraudulent election in a statement released after the certification
      and said he had no regrets, media outlets reported. An Alabama judge denied
      Moore’s request to block certification of the results of the Dec. 12 election in
      a decision shortly before the canvassing board met. Moore’s challenge alleged
      there had been potential voter fraud that denied him a chance of victory. His
      filing on Wednesday in the Montgomery Circuit Court sought to halt the meeting
      scheduled to ratify Jones’ win on Thursday. Moore could ask for a recount, in
      addition to possible other court challenges, Merrill said in an interview with
      Fox News Channel. He would have to complete paperwork "within a timed period"
      and show he has the money for a challenge, Merrill said. "We’ve not been
      notified yet of their intention to do that," Merrill said. Regarding the claim
      of voter fraud, Merrill told CNN that more than 100 cases had been reported.
      "We’ve adjudicated more than 60 of those. We will continue to do that," he said.
      Republican lawmakers in Washington had distanced themselves from Moore and
      called for him to drop out of the race after several women accused him of sexual
      assault or misconduct dating back to when they were teenagers and he was in his
      early 30s.  Moore has denied wrongdoing and Reuters has not been able to
      independently verify the allegations. '
```

Let’s now remove everything except uppercase / lowercase letters using Regular Expressions.

```
[16]: first_text = re.sub('\[[^]]*\]', ' ', first_text)
      first_text = re.sub('[^a-zA-Z]',' ',first_text)  # replaces non-alphabets with␣
        ↪spaces
      first_text = first_text.lower() # Converting from uppercase to lowercase
      first_text
```

```
[16]: 'politicsnews jones certified u s  senate winner despite moore challenge
      reuters    alabama officials on thursday certified democrat doug jones the
      winner of the state s u s  senate race  after a state judge denied a challenge
      by republican roy moore  whose campaign was derailed by accusations of sexual
      misconduct with teenage girls  jones won the vacant seat by about        votes
      or     percentage points  election officials said  that made him the first
```

```
democrat in a quarter of a century to win a senate seat in alabama   the seat
was previously held by republican jeff sessions  who was tapped by u s
president donald trump as attorney general  a state canvassing board composed of
alabama secretary of state john merrill  governor kay ivey and attorney general
steve marshall certified the election results  seating jones will narrow the
republican majority in the senate to    of     seats  in a statement  jones
called his victory  a new chapter  and pledged to work with both parties  moore
declined to concede defeat even after trump urged him to do so  he stood by
claims of a fraudulent election in a statement released after the certification
and said he had no regrets  media outlets reported  an alabama judge denied
moore s request to block certification of the results of the dec     election in
a decision shortly before the canvassing board met  moore s challenge alleged
there had been potential voter fraud that denied him a chance of victory  his
filing on wednesday in the montgomery circuit court sought to halt the meeting
scheduled to ratify jones  win on thursday  moore could ask for a recount  in
addition to possible other court challenges  merrill said in an interview with
fox news channel  he would have to complete paperwork  within a timed period
and show he has the money for a challenge  merrill said   we ve not been
notified yet of their intention to do that   merrill said  regarding the claim
of voter fraud  merrill told cnn that more than     cases had been reported   we
ve adjudicated more than    of those  we will continue to do that   he said
republican lawmakers in washington had distanced themselves from moore and
called for him to drop out of the race after several women accused him of sexual
assault or misconduct dating back to when they were teenagers and he was in his
early   s   moore has denied wrongdoing and reuters has not been able to
independently verify the allegations  '
```

Let's remove stopwords like is,a,the… Which do not offer much insight.

[19]:

```python
nltk.download("stopwords")
from nltk.corpus import stopwords

# we can use tokenizer instead of split
first_text = nltk.word_tokenize(first_text)
```

```
[nltk_data] Downloading package stopwords to /home/djoe/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

[20]:

```python
first_text = [ word for word in first_text if not word in set(stopwords.
  ↪words("english"))]
```

Lemmatization to bring back multiple forms of same word to their common root like 'coming', 'comes' into 'come'.

[23]:

```python
lemma = nltk.WordNetLemmatizer()
first_text = [ lemma.lemmatize(word) for word in first_text]

first_text = " ".join(first_text)
first_text
```

[23]:

```
'politicsnews jones certified u senate winner despite moore challenge reuters
alabama official thursday certified democrat doug jones winner state u senate
race state judge denied challenge republican roy moore whose campaign derailed
accusation sexual misconduct teenage girl jones vacant seat vote percentage
point election official said made first democrat quarter century win senate seat
alabama seat previously held republican jeff session tapped u president donald
trump attorney general state canvassing board composed alabama secretary state
john merrill governor kay ivey attorney general steve marshall certified
election result seating jones narrow republican majority senate seat statement
jones called victory new chapter pledged work party moore declined concede
defeat even trump urged stood claim fraudulent election statement released
certification said regret medium outlet reported alabama judge denied moore
request block certification result dec election decision shortly canvassing
board met moore challenge alleged potential voter fraud denied chance victory
filing wednesday montgomery circuit court sought halt meeting scheduled ratify
jones win thursday moore could ask recount addition possible court challenge
merrill said interview fox news channel would complete paperwork within timed
period show money challenge merrill said notified yet intention merrill said
regarding claim voter fraud merrill told cnn case reported adjudicated continue
said republican lawmaker washington distanced moore called drop race several
woman accused sexual assault misconduct dating back teenager early moore denied
wrongdoing reuters able independently verify allegation'
```

Performing it for all the examples in the data.

[24]:

```python
#Removal of HTML Contents
def remove_html(text):
    soup = BeautifulSoup(text, "html.parser")
    return soup.get_text()

#Removal of Punctuation Marks
def remove_punctuations(text):
    return re.sub('\[[^]]*\]', '', text)

# Removal of Special Characters
def remove_characters(text):
    return re.sub("[^a-zA-Z]"," ",text)

#Removal of stopwords
def remove_stopwords_and_lemmatization(text):
    final_text = []
    text = text.lower()
    text = nltk.word_tokenize(text)

    for word in text:
        if word not in set(stopwords.words('english')):
            lemma = nltk.WordNetLemmatizer()
            word = lemma.lemmatize(word)
            final_text.append(word)
    return " ".join(final_text)

#Total function
def cleaning(text):
    text = remove_html(text)
    text = remove_punctuations(text)
    text = remove_characters(text)
    text = remove_stopwords_and_lemmatization(text)
    return text

#Apply function on text column
data['text']=data['text'].apply(cleaning)
```

[25]:
```
data.head()
```

[25]:
```
                                                text  target
0  politicsnews u budget fight loom republican fl…       1
1  politicsnews u military accept transgender rec…       1
2  politicsnews senior u republican senator let m…       1
3  politicsnews fbi russia probe helped australia…       1
4  politicsnews trump want postal service charge …       1
```

Train Test Split

[26]:
```
X_train, X_test, y_train, y_test = train_test_split(data['text'],␣
 ↪data['target'], random_state=0)
```

Tokenizing

[31]:
```
max_features = 10000
maxlen = 300
```

[32]:
```
tokenizer = text.Tokenizer(num_words=max_features)
tokenizer.fit_on_texts(X_train)
tokenized_train = tokenizer.texts_to_sequences(X_train)
X_train = sequence.pad_sequences(tokenized_train, maxlen=maxlen)
```

[33]:
```
tokenized_test = tokenizer.texts_to_sequences(X_test)
X_test = sequence.pad_sequences(tokenized_test, maxlen=maxlen)
```

[27]:
```
X_train.size, X_test.size
```

[27]:
```
(33673, 11225)
```

[39]:
```
X_train[3363]
```

```
[39]: 'politicsnews senator grill u education secretary proposal slash budget
      washington reuters u education secretary betsy devos faced hostile question
      senate committee tuesday tried win lawmaker president donald trump proposal
      slash department funding percent devos republican narrowly senate approval post
      february strident opposition democrat fellow party member testified senate
      appropriation subcommittee education proposed budget trump submitted congress
      last month trump plan cut billion education department budget would improve
      educational opportunity shift federal role education devos told panel understand
      figure alarming many said however budget refocuses department supporting state
      school district effort provide high quality education student democrat took turn
      asking devos bigger budget line item talking student say could hurt large
      spending cut pointed exchange whether private school receive federal fund would
      agree discriminate student devos would repeat school taking federal money must
      abide u law senator jeff merkley fellow democrat said refusing answer question
      federal law unclear many area possible discrimination right transgendered people
      lawmaker expected alter trump proposed budget voting subcommittee chair
      conservative republican roy blunt said believed congress would approve budget
      proposed significant cut department budget likely untenable blunt said pressing
      specifically preserve fund technical program work study financial aid special
      olympics civil right group democrat say budget would send public dollar private
      company disband school care hurt school poor neighborhood shrink rank teacher
      make harder many afford college devos currently working major transformation
      student loan budget suggests changing income based repayment plan ending loan
      forgiveness worker public sector devos said would clear confusion around loan
      stated aim giving parent choice child education devos republican support charter
      school publicly funded operate independently frequently corporation well subsidy
      help pay private school tuition many republican panel applauded budget proposal
      boost school choice program subcommittee senior democrat patty murray said cut
      highlight way policy priority president trump pushing would hurt student hurt
      community represent clear broken promise worker middle class'
```

```
[ ]:
```