untitled-3

October 26, 2023

1 Phase 04 Development Part II:

In this phase we have included whats the purpose of deep machine learning and as our project deals with detecting fake news through natural language processing.

2 Natural Language Processing:

Natural language processing (NLP) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate speech. It involves processing natural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves. Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

3 How natural language processing major role in detecting fake news:

To identify bogus news,sentiment analysis using NLP can be effective strategy. NLP algorithms can be ascertain the intention and any biases of an author by analysing the emotions displayed in a news story or social media post. Fake news frequently preys on readers' emotions by using strong language or exaggeration.

4 Is Natural Language Processing actually needed?

NLP is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.

5 Terminologies:

6 Fake News:

```
we are defining "fake news" as those news stories that are false: the story itself is fabr
```

However, it's important to acknowledge that "fake news" is a complex and nuanced problem, c

In order to handle these type of unwanted news,we have implemented specific NLP properties:

7 TF-IDF VECTORIZER:

8 TF(TERM FREQUENCY):

In the document, words are present so many times that is called term frequency. In this section, if you get the largest values it means that word is present so many times with respect to other words. when you get word is parts of speech word that means the document is a very nice match.

9 IDF(INVERSE DOCUMENT FREQUENCY):

In a single document, words are present so many times, but also available so many times in another document also which is not relevant. IDF is a proportion of how critical a term is in the whole corpus. collection of word Documents will convert into the matrix which contains TF-IDF features using TfidfVectorizer.

In the previous phase, we have done the loading and preprocessing of data.Here we are c

[8]:
```
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
```

[9]:
```
df_fake = pd.read_csv("Fake.csv")
df_true = pd.read_csv("True.csv")
```

[10]:
```
df_fake.head()
```

[10]:

	title	text	subject	\
0	Donald Trump Sends Out Embarrassing New Year'…	Donald Trump just couldn t wish all Americans …	News	
1	Drunk Bragging Trump Staffer Started Russian …	House Intelligence Committee Chairman Devin Nu…	News	
2	Sheriff David Clarke Becomes An Internet Joke…	On Friday, it was revealed that former Milwauk…	News	
3	Trump Is So Obsessed He Even Has Obama's Name…	On Christmas day, Donald Trump announced that …	News	
4	Pope Francis Just Called Out Donald Trump Dur…			

```
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
```

[11]:
```
df_true.head(5)
```

[11]:
```
                                               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
```

[12]:
```
df_fake["class"] = 0
df_true["class"] = 1
```

[14]:
```
df_fake.shape, df_true.shape
```

[14]:
```
((23481, 5), (21417, 5))
```

[15]:
```
# 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)
```

[16]:
```
df_fake.shape, df_true.shape
```

[16]:
```
((23471, 5), (21407, 5))
```

[17]:
```
df_fake_manual_testing["class"] = 0
df_true_manual_testing["class"] = 1
```

```
C:\Users\CSE_BAY3\AppData\Local\Temp\ipykernel_5544\860779283.py: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-a-copy
  df_fake_manual_testing["class"] = 0
C:\Users\CSE_BAY3\AppData\Local\Temp\ipykernel_5544\860779283.py: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-a-copy
  df_true_manual_testing["class"] = 1
```

[18]:
```
df_fake_manual_testing.head(10)
```

[18]:

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

```
23477  21st Century Wire says It s a familiar theme. …  Middle-east
23478  Patrick Henningsen  21st Century WireRemember …  Middle-east
23479  21st Century Wire says Al Jazeera America will…  Middle-east
23480  21st Century Wire says As 21WIRE predicted in …  Middle-east

                   date  class
23471  January 20, 2016      0
23472  January 19, 2016      0
23473  January 19, 2016      0
23474  January 19, 2016      0
23475  January 18, 2016      0
23476  January 16, 2016      0
23477  January 16, 2016      0
23478  January 15, 2016      0
23479  January 14, 2016      0
23480  January 12, 2016      0
```

[19]:
```
df_true_manual_testing.head(10)
```

[19]:
```
                                                  title  \
21407  Mata Pires, owner of embattled Brazil builder …
21408  U.S., North Korea clash at U.N. forum over nuc…
21409  U.S., North Korea clash at U.N. arms forum on …
21410  Headless torso could belong to submarine journ…
21411  North Korea shipments to Syria chemical arms a…
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  \
21407  SAO PAULO (Reuters) - Cesar Mata Pires, the ow…  worldnews
21408  GENEVA (Reuters) - North Korea and the United …  worldnews
21409  GENEVA (Reuters) - North Korea and the United …  worldnews
21410  COPENHAGEN (Reuters) - Danish police said on T…  worldnews
21411  UNITED NATIONS (Reuters) - Two North Korean sh…  worldnews
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  class
21407  August 22, 2017       1
21408  August 22, 2017       1
21409  August 22, 2017       1
```

```
21410  August 22, 2017       1
21411  August 21, 2017       1
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
```

[20]:
```
df_manual_testing = pd.concat([df_fake_manual_testing,df_true_manual_testing],␣
↪axis = 0)
df_manual_testing.to_csv("manual_testing.csv")
```

[23]:
```
df_merge = pd.concat([df_fake, df_true], axis =0 )
df_merge.head(10)
```

[23]:
```
                                                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…
5   Racist Alabama Cops Brutalize Black Boy While…
6   Fresh Off The Golf Course, Trump Lashes Out A…
7   Trump Said Some INSANELY Racist Stuff Inside …
8   Former CIA Director Slams Trump Over UN Bully…
9   WATCH: Brand-New Pro-Trump Ad Features So Muc…

                                                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
5  The number of cases of cops brutalizing and ki…    News
6  Donald Trump spent a good portion of his day a…    News
7  In the wake of yet another court decision that…    News
8  Many people have raised the alarm regarding th…    News
9  Just when you might have thought we d get a br…    News

                date  class
0  December 31, 2017      0
1  December 31, 2017      0
2  December 30, 2017      0
3  December 29, 2017      0
4  December 25, 2017      0
5  December 25, 2017      0
6  December 23, 2017      0
```

```
7   December 23, 2017      0
8   December 22, 2017      0
9   December 21, 2017      0
```

[24]:
```
df_merge.columns
```

[24]:
```
Index(['title', 'text', 'subject', 'date', 'class'], dtype='object')
```

[25]:
```
df = df_merge.drop(["title", "subject","date"], axis = 1)
```

[27]:
```
df.isnull().sum()
```

[27]:
```
text     0
class    0
dtype: int64
```

[29]:
```
df = df.sample(frac = 1)
```

[30]:
```
df.head()
```

[30]:
```
                                                    text  class
5391   Make no mistake about it: Ted Cruz is horrible…      0
2841   WASHINGTON (Reuters) - U.S. President Donald T…      1
8914   SAN JUAN (Reuters) - Puerto Rico's semi-public…      1
10062  WASHINGTON (Reuters) - A moderate Republican s…      1
19332  Streep s shameful attempt to paint Trump as a …      0
```

[31]:
```
df.reset_index(inplace = True)
df.drop(["index"], axis = 1, inplace = True)
```

[32]:
```
df.columns
```

[32]:
```
Index(['text', 'class'], dtype='object')
```

[33]:
```
df.head()
```

[33]:
```
                                                text  class
0  Make no mistake about it: Ted Cruz is horrible…      0
1  WASHINGTON (Reuters) - U.S. President Donald T…      1
2  SAN JUAN (Reuters) - Puerto Rico's semi-public…      1
3  WASHINGTON (Reuters) - A moderate Republican s…      1
4  Streep s shameful attempt to paint Trump as a …      0
```

[34]:
```
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
```

[35]:
```
df["text"] = df["text"].apply(wordopt)
```

[36]:
```
x = df["text"]
y = df["class"]
```

[37]:
```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

10 A. VECTORIZATION:

Vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which is used to find word predictions, word similarities/semantics.

For curiosity, you surely want to check out this article on ' Why data are represented as vectors in Data Science Problems'.

To make documents' corpora more relatable for computers, they must first be converted into some numerical structure. There are few techniques that are used to achieve this such as 'Bag of Words'.

[38]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)
```

11 B. LOGISTIC REGRESSION:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

[39]:
```
from sklearn.linear_model import LogisticRegression

LR = LogisticRegression()
LR.fit(xv_train,y_train)
```

[39]:
```
LogisticRegression()
```

[40]:
```
pred_lr=LR.predict(xv_test)
```

[42]:
```
LR.score(xv_test, y_test)
```

[42]:
```
0.986541889483066
```

[43]:
```
print(classification_report(y_test, pred_lr))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5849
           1       0.98      0.99      0.99      5371

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

12 C. DECISION TREE CLASSIFIER:

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

[51]:
```
from sklearn.tree import DecisionTreeClassifier

DT = DecisionTreeClassifier()
DT.fit(xv_train, y_train)
```

[51]:
```
DecisionTreeClassifier()
```

[52]:
```
pred_dt = DT.predict(xv_test)
```

[53]:
```
DT.score(xv_test, y_test)
```

[53]:
```
0.9949197860962566
```

[54]:
```
print(classification_report(y_test, pred_dt))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5849
           1       0.99      1.00      0.99      5371

    accuracy                           0.99     11220
```

```
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

13 D. GRADIENT BOOSTING CLASSIFIER:

Gradient Boosting is a popular boosting algorithm in machine learning used for classification and regression tasks. Boosting is one kind of ensemble Learning method which trains the model sequentially and each new model tries to correct the previous model. It combines several weak learners into strong learners.

[ ]:
```
from sklearn.ensemble import GradientBoostingClassifier

GBC = GradientBoostingClassifier(random_state=0)
GBC.fit(xv_train, y_train)
```

[ ]:
```
pred_gbc = GBC.predict(xv_test)
```

[ ]:
```
GBC.score(xv_test, y_test)
```

[ ]:
```
print(classification_report(y_test, pred_gbc))
```

14 E. RANDOM FOREST CLASSIFIER:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

[ ]:
```
from sklearn.ensemble import RandomForestClassifier

RFC = RandomForestClassifier(random_state=0)
RFC.fit(xv_train, y_train)
```

[ ]:
```
pred_rfc = RFC.predict(xv_test)
```

[ ]:
```
RFC.score(xv_test, y_test)
```

[ ]:
```
print(classification_report(y_test, pred_rfc))
```

[ ]:
```
def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Not A Fake News"

def manual_testing(news):
```

```
    testing_news = {"text":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xv_test = vectorization.transform(new_x_test)
    pred_LR = LR.predict(new_xv_test)
    pred_DT = DT.predict(new_xv_test)
    pred_GBC = GBC.predict(new_xv_test)
    pred_RFC = RFC.predict(new_xv_test)

    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC Prediction: {} \nRFC Prediction: {}".
                 format(output_lable(pred_LR[0]),
                 output_lable(pred_DT[0]),
                 output_lable(pred_GBC[0]),
                 output_lable(pred_RFC[0])))
```

[ ]: