Twitter-Sentiment-Analysis

January 25, 2023

1 Twitter Sentiment Analysis

1.0.1 Problem Statement:

The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor. In this problem, we will take you through a hate speech detection model with Machine Learning and Python.

Hate Speech Detection is generally a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. So, for the task of hate speech detection model, we will use the Twitter tweets to identify tweets containing Hate speech.

1.0.2 Business Understanding:

Detection of hate speech in tweets is an important issue for businesses to consider for several reasons.

First, hate speech can be harmful and offensive to individuals and groups, and businesses have a social responsibility to address it. In addition, businesses may face legal and reputational risks if they fail to address hate speech on their platforms.

Second, businesses that operate social media platforms or engage in social media marketing may need to monitor and address hate speech to maintain the trust and loyalty of their users and customers. If a business is perceived as tolerating hate speech, it may face backlash from users and negative media attention.

Finally, businesses may also have a financial incentive to address hate speech, as it can negatively impact the user experience and drive users away from the platform.

To detect hate speech in tweets, businesses may use a combination of automated tools and human moderation. Automated tools may include machine learning algorithms that are trained to identify hate speech based on certain characteristics, such as the use of certain words or phrases. Human moderation may involve a team of moderators who review tweets and take appropriate action, such as deleting the tweet or banning the user.

It's important to note that detecting hate speech can be challenging, as it may involve complex issues of context and intent. It is also important for businesses to consider the potential for false positives and ensure that their approaches to detecting and addressing hate speech are fair and transparent.

1.1 Import Data Analytics Libraries and Twitter Datasets

[1]:
```
# Pandas is necessary for array manipulation and calculation
import pandas as pd
from pandas import DataFrame

# Matplotlib library to create visualizations
import matplotlib.pyplot as plt

# Missingno used for missing data visualization
import missingno as msno

# NumPy necessary for statistical calculations
import numpy as np
np.set_printoptions(threshold=np.inf)

# Change theme of graphs
plt.style.use('fivethirtyeight')

# Export images from Jupyter to PDF
%matplotlib inline

# Import Seaborn to graph distplots and boxplots
import seaborn as sns

from sklearn.linear_model import LinearRegression

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

1.1.1 Import Datasets

[2]:
```
# Import Training and Test dataset
train = pd.read_csv('train_E6oV3lV.csv')
test = pd.read_csv('test_tweets_anuFYb8.csv')

train_df = pd.read_csv('train_E6oV3lV.csv')
test_df = pd.read_csv('test_tweets_anuFYb8.csv')

train.head()
```

[2]:

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s…
1	2	0	@user @user thanks for #lyft credit i can't us…
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in …
4	5	0	factsguide: society now #motivation

1.2 Data Cleaning and Normalisation

1.2.1 View Headers

[3]:
```
test_df.head()
```

[3]:
```
      id                                                tweet
0  31963  #studiolife #aislife #requires #passion #dedic…
1  31964   @user #white #supremacists want everyone to s…
2  31965  safe ways to heal your #acne!!    #altwaystohe…
3  31966  is the hp and the cursed child book up for res…
4  31967    3rd #bihday to my amazing, hilarious #nephew…
```

[4]:
```
train_df.head()
```

[4]:
```
   id  label                                              tweet
0   1      0   @user when a father is dysfunctional and is s…
1   2      0  @user @user thanks for #lyft credit i can't us…
2   3      0                                    bihday your majesty
3   4      0  #model   i love u take with u all the time in …
4   5      0             factsguide: society now    #motivation
```

1.2.2 Verify Data types

[5]:
```
test_df.dtypes
```

[5]:
```
id        int64
tweet    object
dtype: object
```

[6]:
```
train_df.dtypes
```

[6]:
```
id        int64
label     int64
tweet    object
dtype: object
```

1.2.3 Check for null values

[7]:
```
test_df.isnull().sum()
```

[7]:
```
id       0
tweet    0
dtype: int64
```

[8]:
```
train_df.isnull().sum()
```

[8]:
```
id       0
label    0
tweet    0
dtype: int64
```

1.2.4 Show Shape

[9]:
```
test_df.shape
```

[9]:
```
(17197, 2)
```

[10]:
```
train_df.shape
```

[10]:
```
(31962, 3)
```

1.2.5 Remove Duplicates

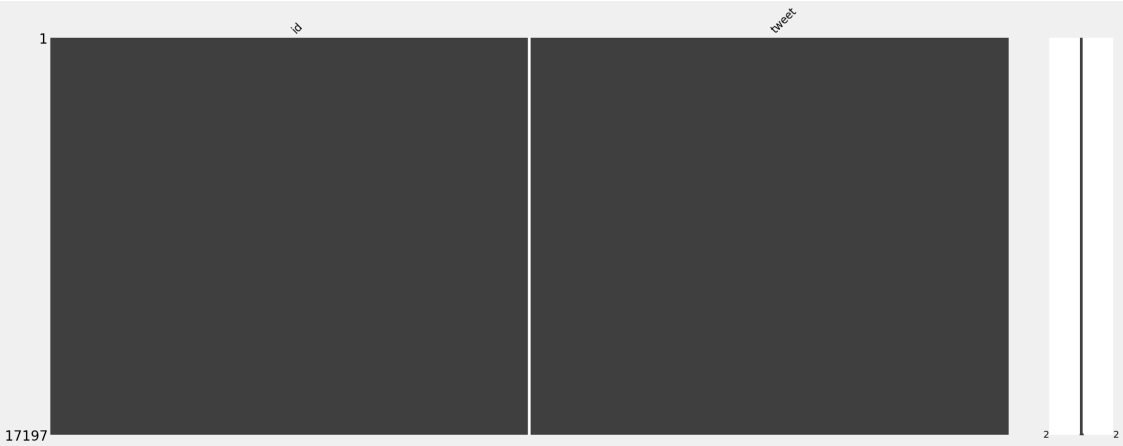
[11]:
```
test_df = test_df.drop_duplicates()
```

[12]:
```
train_df = train_df.drop_duplicates()
```

1.2.6 Check for Missing Data

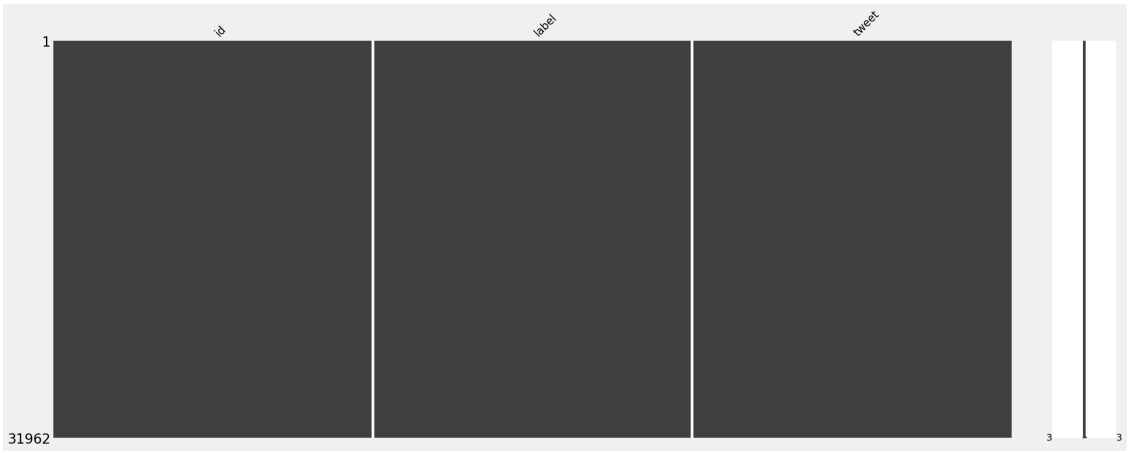
[13]:
```
msno.matrix(test_df)
```

[13]:
```
<AxesSubplot:>
```



[14]:
```
msno.matrix(train_df)
```

[14]:
```
<AxesSubplot:>
```

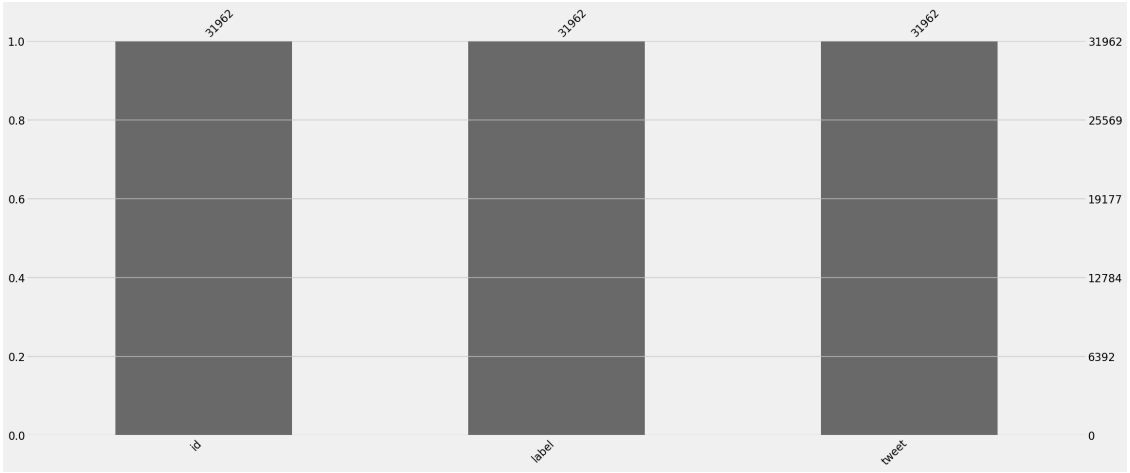


[15]:
```
msno.bar(test_df)
```

[15]: <AxesSubplot:>

[16]:
```
msno.bar(train_df)
```

[16]: <AxesSubplot:>



1.2.7 Remove Punctuation

[17]:
```
# Import string library and define variable containing string characters
import string
punch = string.punctuation
```

[18]:
```
# Function to remove punctuation from specified text
def remove_punctuation(text):
    no_punct=[words for words in text if words not in punch]
    words_wo_punct=''.join(no_punct)
    return words_wo_punct
```

[19]:
```
test_df['tweet'] = test_df['tweet'].apply(lambda x: remove_punctuation(x))
test_df.head()
```

[19]:

	id	tweet
0	31963	studiolife aislife requires passion dedication…
1	31964	user white supremacists want everyone to see …
2	31965	safe ways to heal your acne altwaystoheal h…
3	31966	is the hp and the cursed child book up for res…
4	31967	3rd bihday to my amazing hilarious nephew el…

[20]:
```
train_df['tweet'] = train_df['tweet'].apply(lambda x: remove_punctuation(x))
train_df.head()
```

[20]:

	id	label	tweet
0	1	0	user when a father is dysfunctional and is so…
1	2	0	user user thanks for lyft credit i cant use ca…
2	3	0	bihday your majesty
3	4	0	model i love u take with u all the time in u…
4	5	0	factsguide society now motivation

1.2.8 Tokenization

[21]:
```
import re
```

[22]:
```
def tokenize(text):
    split=re.split("\W+",text)
    return split
```

[23]:
```
test_df['tweet']=test_df['tweet'].apply(lambda x: tokenize(x.lower()))
test_df.head()
```

[23]:

	id	tweet
0	31963	[studiolife, aislife, requires, passion, dedic…
1	31964	[, user, white, supremacists, want, everyone, …
2	31965	[safe, ways, to, heal, your, acne, altwaystohe…
3	31966	[is, the, hp, and, the, cursed, child, book, u…
4	31967	[, 3rd, bihday, to, my, amazing, hilarious, ne…

[24]:
```
train_df['tweet']=train_df['tweet'].apply(lambda x: tokenize(x.lower()))
train_df.head()
```

[24]:

	id	label	tweet
0	1	0	[, user, when, a, father, is, dysfunctional, a…
1	2	0	[user, user, thanks, for, lyft, credit, i, can…
2	3	0	[, bihday, your, majesty]
3	4	0	[model, i, love, u, take, with, u, all, the, t…
4	5	0	[, factsguide, society, now, motivation]

1.2.9 Stop Words

[25]:
```
import nltk

nltk.download('omw-1.4')

# Download stop words (run once)
nltk.download('stopwords')

# Download lemmatizer list (run once)
nltk.download('wordnet')
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
```

```
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[25]:
```
True
```

[26]:
```
stopword = nltk.corpus.stopwords.words('english')
print(stopword[:11])
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've"]
```

[27]:
```
def remove_stopwords(text):
    text=[word for word in text if word not in stopword]
    return text
```

[28]:
```
test_df['tweet'] = test_df['tweet'].apply(lambda x: remove_stopwords(x))
test_df.head()
```

[28]:

	id	tweet
0	31963	[studiolife, aislife, requires, passion, dedic…
1	31964	[, user, white, supremacists, want, everyone, …
2	31965	[safe, ways, heal, acne, altwaystoheal, health…
3	31966	[hp, cursed, child, book, reservations, alread…
4	31967	[, 3rd, bihday, amazing, hilarious, nephew, el…

[29]:
```
train_df['tweet'] = train_df['tweet'].apply(lambda x: remove_stopwords(x))
train_df.head()
```

[29]:

	id	label	tweet
0	1	0	[, user, father, dysfunctional, selfish, drags…
1	2	0	[user, user, thanks, lyft, credit, cant, use, …
2	3	0	[, bihday, majesty]
3	4	0	[model, love, u, take, u, time, urð, ð, ð, ð, …
4	5	0	[, factsguide, society, motivation]

1.2.10 Lemmatization

[30]:
```
w_tokenizer = nltk.tokenize.WhitespaceTokenizer()
lemmatizer = nltk.stem.WordNetLemmatizer()
```

[31]:
```
def lemmatize(s):
     s = [lemmatizer.lemmatize(word) for word in s]
     return s
```

[32]:
```
test_df = test_df.assign(tweet = test_df['tweet'].apply(lambda x: lemmatize(x)))
test_df.head()
```

[32]:

	id	tweet
0	31963	[studiolife, aislife, requires, passion, dedic…
1	31964	[, user, white, supremacist, want, everyone, s…
2	31965	[safe, way, heal, acne, altwaystoheal, healthy…
3	31966	[hp, cursed, child, book, reservation, already…
4	31967	[, 3rd, bihday, amazing, hilarious, nephew, el…

[33]:
```
train_df = train_df.assign(tweet = train_df['tweet'].apply(lambda x:␣
 ↪lemmatize(x)))
train_df.head()
```

[33]:

	id	label	tweet
0	1	0	[, user, father, dysfunctional, selfish, drag,…
1	2	0	[user, user, thanks, lyft, credit, cant, use, …
2	3	0	[, bihday, majesty]
3	4	0	[model, love, u, take, u, time, urð, ð, ð, ð, …
4	5	0	[, factsguide, society, motivation]

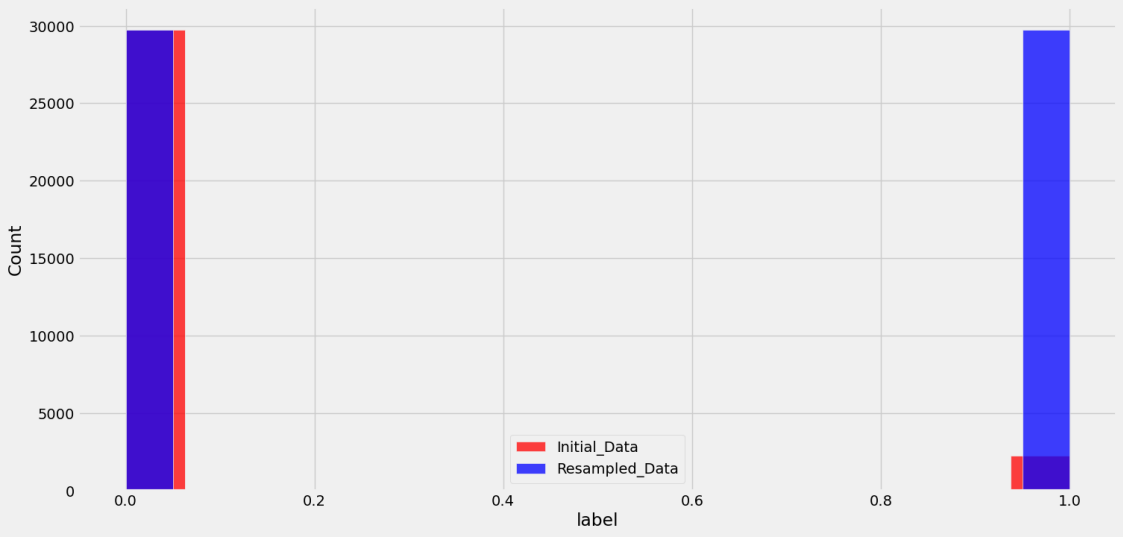
1.3 Representation Learning

[34]:
```
import re
from sklearn.utils import resample

# Clear text of uppercase characters and remove nonstandard text characters
def  clean_text(df, text_field):
    df[text_field] = df[text_field].str.lower()
    df[text_field] = df[text_field].apply(lambda elem: re.
 ↪sub(r"(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)|^rt|http.+?", "",␣
 ↪elem))
    return df
test_clean = clean_text(test, "tweet")
train_clean = clean_text(train, "tweet")

train_majority = train_clean[train_clean.label==0]
train_minority = train_clean[train_clean.label==1]
train_minority_upsampled = resample(train_minority,
                                    replace=True,
                                    n_samples=len(train_majority),
                                    random_state=123)
train_upsampled = pd.concat([train_minority_upsampled, train_majority])
train_upsampled['label'].value_counts()
```

[34]:
```
1    29720
0    29720
Name: label, dtype: int64
```

[35]:
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

fig, axs = plt.subplots(1,2 , figsize=(16,8))
text_pos = " ".join(train_clean['tweet'][train.label == 0])
text_neg = " ".join(train_clean['tweet'][train.label == 1])
train_cloud_pos = WordCloud(collocations = False, background_color = 'white').
 ↪generate(text_pos)
train_cloud_neg = WordCloud(collocations = False, background_color = 'black').
 ↪generate(text_neg)
axs[0].imshow(train_cloud_pos, interpolation='bilinear')
axs[0].axis('off')
axs[0].set_title('Non-Hate Speech Words')
axs[1].imshow(train_cloud_neg, interpolation='bilinear')
axs[1].axis('off')
axs[1].set_title('Hate Speech Words')

plt.show()
```

[36]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Plot Initial and Resampled Data
plt.figure(figsize=(16,8))

sns.histplot(data = train['label'], color='red', legend=True)
sns.histplot(data = train_upsampled['label'], color = 'blue', legend=True)
plt.legend(['Initial_Data', 'Resampled_Data'])
plt.show()
```



[37]:
```
from wordcloud import WordCloud

# Display after Upsampling word clouds

fig, axs = plt.subplots(1,2 , figsize=(16,8))
text_pos = " ".join(train_upsampled['tweet'][train.label == 0])
text_neg = " ".join(train_upsampled['tweet'][train.label == 1])
train_cloud_pos = WordCloud(collocations = False, background_color = 'white').
 ↪generate(text_pos)
train_cloud_neg = WordCloud(collocations = False, background_color = 'black').
 ↪generate(text_neg)
axs[0].imshow(train_cloud_pos, interpolation='bilinear')
axs[0].axis('off')
axs[0].set_title('Non-Hate Speech Words')
axs[1].imshow(train_cloud_neg, interpolation='bilinear')
axs[1].axis('off')
axs[1].set_title('Hate Speech Words')

plt.show()
```

2 Model Creation & Training Model (Richard Flores)

[38]:
```
import xgboost as xgb
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.metrics import f1_score, accuracy_score
from sklearn.model_selection import train_test_split, RepeatedStratifiedKFold,␣
 ↪cross_val_score

# Create XGB Pipeline
pipeline_xgb = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf',  TfidfTransformer()),
    ('nb', xgb.XGBClassifier(use_label =False)),])

x_train, x_test, y_train, y_test = train_test_split(train_upsampled['tweet'],
                                                    ␣
 ↪train_upsampled['label'],random_state = 0)

# Fit Model
model = pipeline_xgb.fit(x_train, y_train)
y_predict = model.predict(x_test)
print('--'* 20)
print('F1 Score: ',f1_score(y_test, y_predict))
print('Accuracy Score: ', accuracy_score(y_test, y_predict))
print('--'*20)
print('---RepeatedKFOLD---')
cv = RepeatedStratifiedKFold(n_splits = 5, n_repeats = 2, random_state =1)
score2 = cross_val_score(pipeline_xgb, x_train, y_train, cv=cv,␣
 ↪scoring='f1_micro', n_jobs=1)
score2 = np.mean(score2)
print('--'* 20)
print('RKFold Score: ', score2)
print('--'* 20)
```

```
[11:02:51] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
```

```
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:02:51] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
----------------------------------------
F1 Score:  0.9159862338889263
Accuracy Score:  0.9162180349932705
----------------------------------------
---RepeatedKFOLD---
[11:02:55] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:02:55] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:02:57] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:02:57] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
```

```
old behavior.
[11:03:00] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:00] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:03] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:03] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:06] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:06] WARNING: C:/Users/Administrator/workspace/xgboost-
```

```
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:09] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:09] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:11] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:11] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:14] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.
```

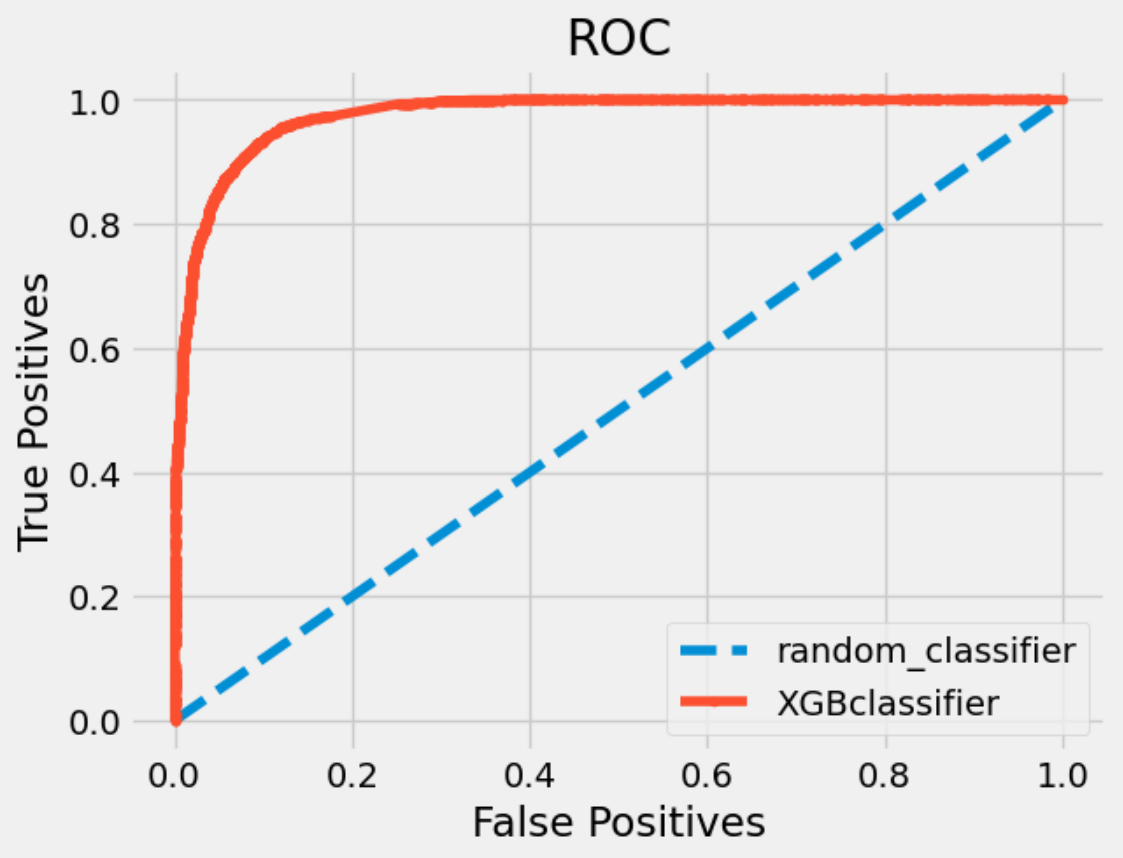
```
[11:03:14] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:17] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:17] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
[11:03:20] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:576:
Parameters: { "use_label" } might not be used.

  This could be a false alarm, with some parameters getting used by language
bindings but
  then being mistakenly passed down to XGBoost core, or some parameter actually
being used
  but getting flagged wrongly here. Please open an issue if you find any such
cases.


[11:03:20] WARNING: C:/Users/Administrator/workspace/xgboost-
win64_release_1.5.1/src/learner.cc:1115: Starting in XGBoost 1.3.0, the default
evaluation metric used with the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if you'd like to restore the
old behavior.
----------------------------------------
RKFold Score:  0.9057424854194706
----------------------------------------
```

[39]:
```
from sklearn.metrics import confusion_matrix, precision_score, recall_score,␣
↪f1_score, roc_auc_score, roc_curve, auc, precision_recall_curve
```

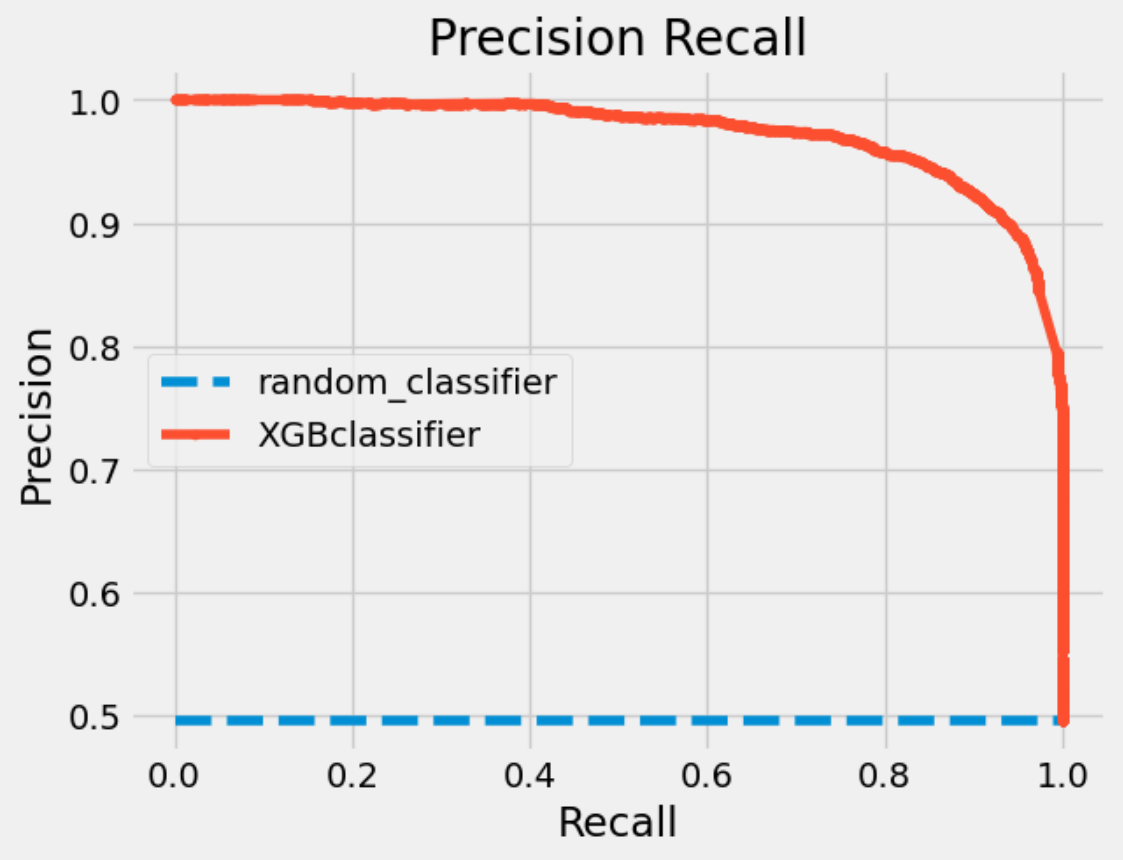
```
print('--- Axis1 = ref_val, Axis0 = pred_val ---')

x_val, y_val = x_test, y_test

yhat = model.predict_proba(x_val)
ypred = model.predict(x_val)
print('Confusion Matrix: ')
print(confusion_matrix(y_val, ypred, labels = [1,0]).T)
print('Precision Score: ', precision_score(y_val, ypred, labels = [1,0]))
print('Recall Score: ', recall_score(y_val, ypred, labels = [1,0]))

print('y_hat_shape: ',yhat.shape)
yhat = yhat[:,1]

ns_probs = [0 for _ in range(len(y_val))]
ns_auc = roc_auc_score(y_val, ns_probs, labels = [1,0])
lr_auc = roc_auc_score(y_val, yhat, labels = [1,0])
print()
print('random_classifier: ROC AUC=%.3f' % (ns_auc))
print('XGBclassifier: ROC AUC=%.3f' % (lr_auc))

ns_fpr, ns_tpr, _ = roc_curve(y_val, ns_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_val, yhat)

plt.plot(ns_fpr, ns_tpr, linestyle='--', label='random_classifier')
plt.plot(lr_fpr, lr_tpr, marker='.', label='XGBclassifier')
plt.xlabel('False Positives')
plt.ylabel('True Positives')
plt.legend()
plt.title('ROC')
plt.show()

lr_precision, lr_recall, _ = precision_recall_curve(y_val, yhat)
print('auc-score: ',auc(lr_recall, lr_precision))
print('f1-score: ', f1_score(y_val, ypred, labels = [1,0]))
no_skill = len(y_val[y_val==1]) / len(y_val)
plt.plot([0, 1], [no_skill, no_skill], linestyle='--',␣
 ↪label='random_classifier')
plt.plot(lr_recall, lr_precision, marker='.', label='XGBclassifier')
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.legend()
plt.title('Precision Recall')
plt.show()
```

```
--- Axis1 = ref_val, Axis0 = pred_val ---
Confusion Matrix:
```

```
[[6787  662]
 [ 583 6828]]
Precision Score:  0.9111290106054504
Recall Score:  0.9208955223880597
y_hat_shape:  (14860, 2)

random_classifier: ROC AUC=0.500
XGBclassifier: ROC AUC=0.976
```




```
auc-score:  0.9733875328219913
f1-score:  0.9159862338889263
```



3 Model Creation & Training Model (Christos Christoforou)

[40]:
```
pip install simpletransformers
```

```
Collecting simpletransformers
  Using cached simpletransformers-0.63.9-py3-none-any.whl (250 kB)
Collecting transformers>=4.6.0
  Using cached transformers-4.26.0-py3-none-any.whl (6.3 MB)
Requirement already satisfied: tqdm>=4.47.0 in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (4.64.1)
Requirement already satisfied: scipy in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.6.2)
Collecting seqeval
  Using cached seqeval-1.2.2.tar.gz (43 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Requirement already satisfied: pandas in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.4.4)
Requirement already satisfied: regex in c:\users\richard\anaconda3\lib\site-
```

```
packages (from simpletransformers) (2022.7.9)
Collecting wandb>=0.10.32
  Using cached wandb-0.13.9-py2.py3-none-any.whl (2.0 MB)
Requirement already satisfied: scikit-learn in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (1.1.3)
Collecting sentencepiece
  Using cached sentencepiece-0.1.97-cp38-cp38-win_amd64.whl (1.1 MB)
Collecting streamlit
  Using cached streamlit-1.17.0-py2.py3-none-any.whl (9.3 MB)
Requirement already satisfied: requests in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (2.28.1)
Collecting tokenizers
  Using cached tokenizers-0.13.2-cp38-cp38-win_amd64.whl (3.3 MB)
Requirement already satisfied: tensorboard in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (2.8.0)
Requirement already satisfied: numpy in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.22.2)
Collecting datasets
  Using cached datasets-2.8.0-py3-none-any.whl (452 kB)
Requirement already satisfied: colorama in c:\users\richard\anaconda3\lib\site-
packages (from tqdm>=4.47.0->simpletransformers) (0.4.5)
Requirement already satisfied: filelock in c:\users\richard\anaconda3\lib\site-
packages (from transformers>=4.6.0->simpletransformers) (3.6.0)
Requirement already satisfied: pyyaml>=5.1 in
c:\users\richard\anaconda3\lib\site-packages (from
transformers>=4.6.0->simpletransformers) (6.0)
Requirement already satisfied: packaging>=20.0 in
c:\users\richard\anaconda3\lib\site-packages (from
transformers>=4.6.0->simpletransformers) (21.3)
Collecting huggingface-hub<1.0,>=0.11.0
  Using cached huggingface_hub-0.11.1-py3-none-any.whl (182 kB)
Requirement already satisfied: setuptools in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (65.5.0)
Collecting docker-pycreds>=0.4.0
  Using cached docker_pycreds-0.4.0-py2.py3-none-any.whl (9.0 kB)
Collecting appdirs>=1.4.3
  Using cached appdirs-1.4.4-py2.py3-none-any.whl (9.6 kB)
Collecting setproctitle
  Using cached setproctitle-1.3.2-cp38-cp38-win_amd64.whl (11 kB)
Collecting GitPython>=1.0.0
  Using cached GitPython-3.1.30-py3-none-any.whl (184 kB)
Requirement already satisfied: pathtools in c:\users\richard\anaconda3\lib\site-
packages (from wandb>=0.10.32->simpletransformers) (0.1.2)
Requirement already satisfied: psutil>=5.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (5.9.0)
Requirement already satisfied: Click!=8.0.0,>=7.0 in
```

```
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (8.0.4)
Requirement already satisfied: typing-extensions in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (4.3.0)
Collecting sentry-sdk>=1.0.0
  Using cached sentry_sdk-1.14.0-py2.py3-none-any.whl (178 kB)
Requirement already satisfied: protobuf!=4.21.0,<5,>=3.19.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (3.19.4)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(2022.9.24)
Requirement already satisfied: charset-normalizer<3,>=2 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(2.0.4)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(3.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(1.26.12)
Collecting pyarrow>=6.0.0
  Downloading pyarrow-10.0.1-cp38-cp38-win_amd64.whl (20.3 MB)
     ---------------------------------------- 20.3/20.3 MB 9.5 MB/s eta 0:00:00
Collecting multiprocess
  Downloading multiprocess-0.70.14-py38-none-any.whl (132 kB)
     ---------------------------------------- 132.0/132.0 kB 1.6 MB/s eta 0:00:00
Collecting aiohttp
  Downloading aiohttp-3.8.3-cp38-cp38-win_amd64.whl (324 kB)
     ---------------------------------------- 324.3/324.3 kB 2.5 MB/s eta 0:00:00
Requirement already satisfied: dill<0.3.7 in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(0.3.4)
Collecting responses<0.19
  Downloading responses-0.18.0-py3-none-any.whl (38 kB)
Collecting xxhash
  Downloading xxhash-3.2.0-cp38-cp38-win_amd64.whl (30 kB)
Requirement already satisfied: fsspec[http]>=2021.11.1 in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(2022.10.0)
Requirement already satisfied: pytz>=2020.1 in
c:\users\richard\anaconda3\lib\site-packages (from pandas->simpletransformers)
(2022.1)
Requirement already satisfied: python-dateutil>=2.8.1 in
c:\users\richard\anaconda3\lib\site-packages (from pandas->simpletransformers)
(2.8.2)
Requirement already satisfied: joblib>=1.0.0 in
```

```
c:\users\richard\anaconda3\lib\site-packages (from scikit-
learn->simpletransformers) (1.1.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from scikit-
learn->simpletransformers) (2.2.0)
Collecting tzlocal>=1.1
  Downloading tzlocal-4.2-py3-none-any.whl (19 kB)
Requirement already satisfied: pillow>=6.2.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (9.2.0)
Collecting blinker>=1.0.0
  Downloading blinker-1.5-py2.py3-none-any.whl (12 kB)
Collecting altair>=3.2.0
  Downloading altair-4.2.0-py3-none-any.whl (812 kB)
     ------------------------------------ 812.8/812.8 kB 934.3 kB/s eta 0:00:00
Requirement already satisfied: tornado>=5.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (6.1)
Requirement already satisfied: watchdog in c:\users\richard\anaconda3\lib\site-
packages (from streamlit->simpletransformers) (1.0.2)
Collecting validators>=0.2
  Downloading validators-0.20.0.tar.gz (30 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Collecting rich>=10.11.0
  Downloading rich-13.2.0-py3-none-any.whl (238 kB)
     ------------------------------------- 238.9/238.9 kB 1.8 MB/s eta 0:00:00
Collecting pydeck>=0.1.dev5
  Downloading pydeck-0.8.0-py2.py3-none-any.whl (4.7 MB)
     ---------------------------------------- 4.7/4.7 MB 7.0 MB/s eta 0:00:00
Collecting semver
  Downloading semver-2.13.0-py2.py3-none-any.whl (12 kB)
Requirement already satisfied: cachetools>=4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (5.0.0)
Requirement already satisfied: toml in c:\users\richard\anaconda3\lib\site-
packages (from streamlit->simpletransformers) (0.10.2)
Requirement already satisfied: importlib-metadata>=1.4 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (4.11.3)
Collecting pympler>=0.9
  Downloading Pympler-1.0.1-py3-none-any.whl (164 kB)
     ------------------------------------- 164.8/164.8 kB 3.3 MB/s eta 0:00:00
Requirement already satisfied: tensorboard-data-server<0.7.0,>=0.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (0.6.1)
Requirement already satisfied: google-auth<3,>=1.6.3 in
c:\users\richard\anaconda3\lib\site-packages (from
```

```
tensorboard->simpletransformers) (2.6.0)
Requirement already satisfied: wheel>=0.26 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (0.37.1)
Requirement already satisfied: absl-py>=0.4 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (1.0.0)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (1.8.1)
Requirement already satisfied: grpcio>=1.24.3 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (1.43.0)
Requirement already satisfied: google-auth-oauthlib<0.5,>=0.4.1 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (0.4.6)
Requirement already satisfied: markdown>=2.6.8 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (3.3.6)
Requirement already satisfied: werkzeug>=0.11.15 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (2.0.3)
Requirement already satisfied: six in c:\users\richard\anaconda3\lib\site-
packages (from absl-py>=0.4->tensorboard->simpletransformers) (1.16.0)
Requirement already satisfied: entrypoints in
c:\users\richard\anaconda3\lib\site-packages (from
altair>=3.2.0->streamlit->simpletransformers) (0.4)
Requirement already satisfied: jinja2 in c:\users\richard\anaconda3\lib\site-
packages (from altair>=3.2.0->streamlit->simpletransformers) (3.1.2)
Requirement already satisfied: toolz in c:\users\richard\anaconda3\lib\site-
packages (from altair>=3.2.0->streamlit->simpletransformers) (0.12.0)
Requirement already satisfied: jsonschema>=3.0 in
c:\users\richard\anaconda3\lib\site-packages (from
altair>=3.2.0->streamlit->simpletransformers) (4.16.0)
Collecting async-timeout<5.0,>=4.0.0a3
  Downloading async_timeout-4.0.2-py3-none-any.whl (5.8 kB)
Collecting aiosignal>=1.1.2
  Downloading aiosignal-1.3.1-py3-none-any.whl (7.6 kB)
Collecting yarl<2.0,>=1.0
  Downloading yarl-1.8.2-cp38-cp38-win_amd64.whl (56 kB)
     ------------------------------------ 56.9/56.9 kB 148.9 kB/s eta 0:00:00
Collecting frozenlist>=1.1.1
  Downloading frozenlist-1.3.3-cp38-cp38-win_amd64.whl (34 kB)
Requirement already satisfied: attrs>=17.3.0 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (21.4.0)
Collecting multidict<7.0,>=4.5
  Downloading multidict-6.0.4-cp38-cp38-win_amd64.whl (28 kB)
```

```
Collecting gitdb<5,>=4.0.1
  Downloading gitdb-4.0.10-py3-none-any.whl (62 kB)
     ---------------------------------------- 62.7/62.7 kB 558.5 kB/s eta 0:00:00
Requirement already satisfied: pyasn1-modules>=0.2.1 in
c:\users\richard\anaconda3\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard->simpletransformers) (0.2.8)
Requirement already satisfied: rsa<5,>=3.1.4 in
c:\users\richard\anaconda3\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard->simpletransformers) (4.8)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
c:\users\richard\anaconda3\lib\site-packages (from google-auth-
oauthlib<0.5,>=0.4.1->tensorboard->simpletransformers) (1.3.1)
Requirement already satisfied: zipp>=0.5 in c:\users\richard\anaconda3\lib\site-
packages (from importlib-metadata>=1.4->streamlit->simpletransformers) (3.8.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
c:\users\richard\anaconda3\lib\site-packages (from
packaging>=20.0->transformers>=4.6.0->simpletransformers) (3.0.9)
Requirement already satisfied: pygments<3.0.0,>=2.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from
rich>=10.11.0->streamlit->simpletransformers) (2.11.2)
Collecting markdown-it-py<3.0.0,>=2.1.0
  Downloading markdown_it_py-2.1.0-py3-none-any.whl (84 kB)
     ---------------------------------------- 84.5/84.5 kB 594.4 kB/s eta 0:00:00
Collecting tzdata
  Downloading tzdata-2022.7-py2.py3-none-any.whl (340 kB)
     ---------------------------------------- 340.1/340.1 kB 843.8 kB/s eta 0:00:00
Collecting pytz-deprecation-shim
  Downloading pytz_deprecation_shim-0.1.0.post0-py2.py3-none-any.whl (15 kB)
Collecting backports.zoneinfo
  Downloading backports.zoneinfo-0.2.1-cp38-cp38-win_amd64.whl (38 kB)
Requirement already satisfied: decorator>=3.4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
validators>=0.2->streamlit->simpletransformers) (5.1.1)
Collecting dill<0.3.7
  Downloading dill-0.3.6-py3-none-any.whl (110 kB)
     ---------------------------------------- 110.5/110.5 kB 1.3 MB/s eta 0:00:00
Collecting smmap<6,>=3.0.1
  Downloading smmap-5.0.0-py3-none-any.whl (24 kB)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\users\richard\anaconda3\lib\site-packages (from
jinja2->altair>=3.2.0->streamlit->simpletransformers) (2.1.1)
Requirement already satisfied: importlib-resources>=1.4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
jsonschema>=3.0->altair>=3.2.0->streamlit->simpletransformers) (5.2.0)
Requirement already satisfied: pkgutil-resolve-name>=1.3.10 in
c:\users\richard\anaconda3\lib\site-packages (from
jsonschema>=3.0->altair>=3.2.0->streamlit->simpletransformers) (1.3.10)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in
```

```
c:\users\richard\anaconda3\lib\site-packages (from
jsonschema>=3.0->altair>=3.2.0->streamlit->simpletransformers) (0.18.0)
Collecting mdurl~=0.1
  Downloading mdurl-0.1.2-py3-none-any.whl (10.0 kB)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
c:\users\richard\anaconda3\lib\site-packages (from
pyasn1-modules>=0.2.1->google-auth<3,>=1.6.3->tensorboard->simpletransformers)
(0.4.8)
Requirement already satisfied: oauthlib>=3.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from requests-
oauthlib>=0.7.0->google-auth-
oauthlib<0.5,>=0.4.1->tensorboard->simpletransformers) (3.2.0)
Building wheels for collected packages: seqeval, validators
  Building wheel for seqeval (setup.py): started
  Building wheel for seqeval (setup.py): finished with status 'done'
  Created wheel for seqeval: filename=seqeval-1.2.2-py3-none-any.whl size=16165
sha256=73d24b53f63b202c1547e160f1c916c761ac34a57f5d64c99118adc557e27cd3
  Stored in directory: c:\users\richard\appdata\local\pip\cache\wheels\ad\5c\ba\
05fa33fa5855777b7d686e843ec07452f22a66a138e290e732
  Building wheel for validators (setup.py): started
  Building wheel for validators (setup.py): finished with status 'done'
  Created wheel for validators: filename=validators-0.20.0-py3-none-any.whl
size=19579
sha256=81fbd04d75440ac249c5ba2ee545dc150db18a480e876c915e78c30e20bbba0d
  Stored in directory: c:\users\richard\appdata\local\pip\cache\wheels\19\09\72\
3eb74d236bb48bd0f3c6c3c83e4e0c5bbfcbcad7c6c3539db8
Successfully built seqeval validators
Installing collected packages: tokenizers, sentencepiece, appdirs, xxhash,
validators, tzdata, smmap, setproctitle, sentry-sdk, semver, pympler, pyarrow,
multidict, mdurl, frozenlist, docker-pycreds, dill, blinker, backports.zoneinfo,
async-timeout, yarl, responses, pytz-deprecation-shim, pydeck, multiprocess,
markdown-it-py, huggingface-hub, gitdb, aiosignal, tzlocal, transformers,
seqeval, rich, GitPython, altair, aiohttp, wandb, streamlit, datasets,
simpletransformers
  Attempting uninstall: dill
    Found existing installation: dill 0.3.4
    Uninstalling dill-0.3.4:
      Successfully uninstalled dill-0.3.4
  Attempting uninstall: huggingface-hub
    Found existing installation: huggingface-hub 0.10.1
    Uninstalling huggingface-hub-0.10.1:
      Successfully uninstalled huggingface-hub-0.10.1
Successfully installed GitPython-3.1.30 aiohttp-3.8.3 aiosignal-1.3.1
altair-4.2.0 appdirs-1.4.4 async-timeout-4.0.2 backports.zoneinfo-0.2.1
blinker-1.5 datasets-2.8.0 dill-0.3.6 docker-pycreds-0.4.0 frozenlist-1.3.3
gitdb-4.0.10 huggingface-hub-0.11.1 markdown-it-py-2.1.0 mdurl-0.1.2
multidict-6.0.4 multiprocess-0.70.14 pyarrow-10.0.1 pydeck-0.8.0 pympler-1.0.1
pytz-deprecation-shim-0.1.0.post0 responses-0.18.0 rich-13.2.0 semver-2.13.0
```

```
sentencepiece-0.1.97 sentry-sdk-1.14.0 seqeval-1.2.2 setproctitle-1.3.2
simpletransformers-0.63.9 smmap-5.0.0 streamlit-1.17.0 tokenizers-0.13.2
transformers-4.26.0 tzdata-2022.7 tzlocal-4.2 validators-0.20.0 wandb-0.13.9
xxhash-3.2.0 yarl-1.8.2
Note: you may need to restart the kernel to use updated packages.
```

[41]:
```
import numpy as np
import pandas as pd

import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet
from nltk.corpus import stopwords
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('punkt')

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
import torch
from torch import nn, optim
from torch.utils.data import TensorDataset, DataLoader

!pip install simpletransformers
from simpletransformers.classification import ClassificationModel,␣
 ↪ClassificationArgs
from transformers import RobertaForSequenceClassification

from wordcloud import WordCloud
from wordcloud import STOPWORDS
import matplotlib.pyplot as plt

## ignore warnings
import warnings
warnings.filterwarnings("ignore",module = "matplotlib\..*")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
```

```
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Richard\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!

Requirement already satisfied: simpletransformers in
c:\users\richard\anaconda3\lib\site-packages (0.63.9)
Requirement already satisfied: requests in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (2.28.1)
Requirement already satisfied: seqeval in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.2.2)
Requirement already satisfied: wandb>=0.10.32 in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (0.13.9)
Requirement already satisfied: regex in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (2022.7.9)
Requirement already satisfied: tokenizers in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (0.13.2)
Requirement already satisfied: scipy in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.6.2)
Requirement already satisfied: scikit-learn in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (1.1.3)
Requirement already satisfied: transformers>=4.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (4.26.0)
Requirement already satisfied: datasets in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (2.8.0)
Requirement already satisfied: tqdm>=4.47.0 in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (4.64.1)
Requirement already satisfied: tensorboard in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (2.8.0)
Requirement already satisfied: numpy in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.22.2)
Requirement already satisfied: pandas in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.4.4)
Requirement already satisfied: sentencepiece in
c:\users\richard\anaconda3\lib\site-packages (from simpletransformers) (0.1.97)
Requirement already satisfied: streamlit in c:\users\richard\anaconda3\lib\site-
packages (from simpletransformers) (1.17.0)
Requirement already satisfied: colorama in c:\users\richard\anaconda3\lib\site-
packages (from tqdm>=4.47.0->simpletransformers) (0.4.5)
Requirement already satisfied: pyyaml>=5.1 in
c:\users\richard\anaconda3\lib\site-packages (from
transformers>=4.6.0->simpletransformers) (6.0)
Requirement already satisfied: filelock in c:\users\richard\anaconda3\lib\site-
packages (from transformers>=4.6.0->simpletransformers) (3.6.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.11.0 in
c:\users\richard\anaconda3\lib\site-packages (from
transformers>=4.6.0->simpletransformers) (0.11.1)
Requirement already satisfied: packaging>=20.0 in
```

```
c:\users\richard\anaconda3\lib\site-packages (from
transformers>=4.6.0->simpletransformers) (21.3)
Requirement already satisfied: Click!=8.0.0,>=7.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (8.0.4)
Requirement already satisfied: appdirs>=1.4.3 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (1.4.4)
Requirement already satisfied: pathtools in c:\users\richard\anaconda3\lib\site-
packages (from wandb>=0.10.32->simpletransformers) (0.1.2)
Requirement already satisfied: setuptools in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (65.5.0)
Requirement already satisfied: docker-pycreds>=0.4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (0.4.0)
Requirement already satisfied: sentry-sdk>=1.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (1.14.0)
Requirement already satisfied: psutil>=5.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (5.9.0)
Requirement already satisfied: typing-extensions in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (4.3.0)
Requirement already satisfied: protobuf!=4.21.0,<5,>=3.19.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (3.19.4)
Requirement already satisfied: setproctitle in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (1.3.2)
Requirement already satisfied: GitPython>=1.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from
wandb>=0.10.32->simpletransformers) (3.1.30)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(1.26.12)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(2022.9.24)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(3.4)
Requirement already satisfied: charset-normalizer<3,>=2 in
c:\users\richard\anaconda3\lib\site-packages (from requests->simpletransformers)
(2.0.4)
Requirement already satisfied: xxhash in c:\users\richard\anaconda3\lib\site-
packages (from datasets->simpletransformers) (3.2.0)
```

```
Requirement already satisfied: fsspec[http]>=2021.11.1 in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(2022.10.0)
Requirement already satisfied: dill<0.3.7 in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(0.3.6)
Requirement already satisfied: multiprocess in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(0.70.14)
Requirement already satisfied: pyarrow>=6.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(10.0.1)
Requirement already satisfied: aiohttp in c:\users\richard\anaconda3\lib\site-
packages (from datasets->simpletransformers) (3.8.3)
Requirement already satisfied: responses<0.19 in
c:\users\richard\anaconda3\lib\site-packages (from datasets->simpletransformers)
(0.18.0)
Requirement already satisfied: python-dateutil>=2.8.1 in
c:\users\richard\anaconda3\lib\site-packages (from pandas->simpletransformers)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\users\richard\anaconda3\lib\site-packages (from pandas->simpletransformers)
(2022.1)
Requirement already satisfied: joblib>=1.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from scikit-
learn->simpletransformers) (1.1.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from scikit-
learn->simpletransformers) (2.2.0)
Requirement already satisfied: blinker>=1.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (1.5)
Requirement already satisfied: watchdog in c:\users\richard\anaconda3\lib\site-
packages (from streamlit->simpletransformers) (1.0.2)
Requirement already satisfied: altair>=3.2.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (4.2.0)
Requirement already satisfied: toml in c:\users\richard\anaconda3\lib\site-
packages (from streamlit->simpletransformers) (0.10.2)
Requirement already satisfied: rich>=10.11.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (13.2.0)
Requirement already satisfied: tornado>=5.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (6.1)
Requirement already satisfied: cachetools>=4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (5.0.0)
```

```
Requirement already satisfied: pympler>=0.9 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (1.0.1)
Requirement already satisfied: semver in c:\users\richard\anaconda3\lib\site-
packages (from streamlit->simpletransformers) (2.13.0)
Requirement already satisfied: tzlocal>=1.1 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (4.2)
Requirement already satisfied: pydeck>=0.1.dev5 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (0.8.0)
Requirement already satisfied: importlib-metadata>=1.4 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (4.11.3)
Requirement already satisfied: pillow>=6.2.0 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (9.2.0)
Requirement already satisfied: validators>=0.2 in
c:\users\richard\anaconda3\lib\site-packages (from
streamlit->simpletransformers) (0.20.0)
Requirement already satisfied: markdown>=2.6.8 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (3.3.6)
Requirement already satisfied: wheel>=0.26 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (0.37.1)
Requirement already satisfied: grpcio>=1.24.3 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (1.43.0)
Requirement already satisfied: google-auth-oauthlib<0.5,>=0.4.1 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (0.4.6)
Requirement already satisfied: tensorboard-data-server<0.7.0,>=0.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (0.6.1)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (1.8.1)
Requirement already satisfied: werkzeug>=0.11.15 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (2.0.3)
Requirement already satisfied: absl-py>=0.4 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (1.0.0)
Requirement already satisfied: google-auth<3,>=1.6.3 in
c:\users\richard\anaconda3\lib\site-packages (from
tensorboard->simpletransformers) (2.6.0)
Requirement already satisfied: six in c:\users\richard\anaconda3\lib\site-
```

```
packages (from absl-py>=0.4->tensorboard->simpletransformers) (1.16.0)
Requirement already satisfied: jinja2 in c:\users\richard\anaconda3\lib\site-
packages (from altair>=3.2.0->streamlit->simpletransformers) (3.1.2)
Requirement already satisfied: toolz in c:\users\richard\anaconda3\lib\site-
packages (from altair>=3.2.0->streamlit->simpletransformers) (0.12.0)
Requirement already satisfied: jsonschema>=3.0 in
c:\users\richard\anaconda3\lib\site-packages (from
altair>=3.2.0->streamlit->simpletransformers) (4.16.0)
Requirement already satisfied: entrypoints in
c:\users\richard\anaconda3\lib\site-packages (from
altair>=3.2.0->streamlit->simpletransformers) (0.4)
Requirement already satisfied: multidict<7.0,>=4.5 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (6.0.4)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (4.0.2)
Requirement already satisfied: yarl<2.0,>=1.0 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (1.8.2)
Requirement already satisfied: attrs>=17.3.0 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (21.4.0)
Requirement already satisfied: frozenlist>=1.1.1 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (1.3.3)
Requirement already satisfied: aiosignal>=1.1.2 in
c:\users\richard\anaconda3\lib\site-packages (from
aiohttp->datasets->simpletransformers) (1.3.1)
Requirement already satisfied: gitdb<5,>=4.0.1 in
c:\users\richard\anaconda3\lib\site-packages (from
GitPython>=1.0.0->wandb>=0.10.32->simpletransformers) (4.0.10)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
c:\users\richard\anaconda3\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard->simpletransformers) (0.2.8)
Requirement already satisfied: rsa<5,>=3.1.4 in
c:\users\richard\anaconda3\lib\site-packages (from google-
auth<3,>=1.6.3->tensorboard->simpletransformers) (4.8)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
c:\users\richard\anaconda3\lib\site-packages (from google-auth-
oauthlib<0.5,>=0.4.1->tensorboard->simpletransformers) (1.3.1)
Requirement already satisfied: zipp>=0.5 in c:\users\richard\anaconda3\lib\site-
packages (from importlib-metadata>=1.4->streamlit->simpletransformers) (3.8.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
c:\users\richard\anaconda3\lib\site-packages (from
packaging>=20.0->transformers>=4.6.0->simpletransformers) (3.0.9)
Requirement already satisfied: markdown-it-py<3.0.0,>=2.1.0 in
c:\users\richard\anaconda3\lib\site-packages (from
```

```
rich>=10.11.0->streamlit->simpletransformers) (2.1.0)
Requirement already satisfied: pygments<3.0.0,>=2.6.0 in
c:\users\richard\anaconda3\lib\site-packages (from
rich>=10.11.0->streamlit->simpletransformers) (2.11.2)
Requirement already satisfied: tzdata in c:\users\richard\anaconda3\lib\site-
packages (from tzlocal>=1.1->streamlit->simpletransformers) (2022.7)
Requirement already satisfied: backports.zoneinfo in
c:\users\richard\anaconda3\lib\site-packages (from
tzlocal>=1.1->streamlit->simpletransformers) (0.2.1)
Requirement already satisfied: pytz-deprecation-shim in
c:\users\richard\anaconda3\lib\site-packages (from
tzlocal>=1.1->streamlit->simpletransformers) (0.1.0.post0)
Requirement already satisfied: decorator>=3.4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
validators>=0.2->streamlit->simpletransformers) (5.1.1)
Requirement already satisfied: smmap<6,>=3.0.1 in
c:\users\richard\anaconda3\lib\site-packages (from
gitdb<5,>=4.0.1->GitPython>=1.0.0->wandb>=0.10.32->simpletransformers) (5.0.0)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\users\richard\anaconda3\lib\site-packages (from
jinja2->altair>=3.2.0->streamlit->simpletransformers) (2.1.1)
Requirement already satisfied: pkgutil-resolve-name>=1.3.10 in
c:\users\richard\anaconda3\lib\site-packages (from
jsonschema>=3.0->altair>=3.2.0->streamlit->simpletransformers) (1.3.10)
Requirement already satisfied: importlib-resources>=1.4.0 in
c:\users\richard\anaconda3\lib\site-packages (from
jsonschema>=3.0->altair>=3.2.0->streamlit->simpletransformers) (5.2.0)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in
c:\users\richard\anaconda3\lib\site-packages (from
jsonschema>=3.0->altair>=3.2.0->streamlit->simpletransformers) (0.18.0)
Requirement already satisfied: mdurl~=0.1 in
c:\users\richard\anaconda3\lib\site-packages (from markdown-it-
py<3.0.0,>=2.1.0->rich>=10.11.0->streamlit->simpletransformers) (0.1.2)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
c:\users\richard\anaconda3\lib\site-packages (from
pyasn1-modules>=0.2.1->google-auth<3,>=1.6.3->tensorboard->simpletransformers)
(0.4.8)
Requirement already satisfied: oauthlib>=3.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from requests-
oauthlib>=0.7.0->google-auth-
oauthlib<0.5,>=0.4.1->tensorboard->simpletransformers) (3.2.0)
```

[42]:
```
train_data = pd.read_csv('train_E6oV3lV.csv')
test_data = pd.read_csv('test_tweets_anuFYb8.csv')
```

[44]:
```
## drop the id column
train_data.drop('id', axis=1, inplace=True)
```

[45]:
```
train_data.drop_duplicates(inplace=True)
```

[47]:
```
pip install imblearn
```

```
Collecting imblearn
  Downloading imblearn-0.0-py2.py3-none-any.whl (1.9 kB)
Collecting imbalanced-learn
  Downloading imbalanced_learn-0.10.1-py3-none-any.whl (226 kB)
     ---------------------------------- 226.0/226.0 kB 345.2 kB/s eta 0:00:00
Requirement already satisfied: scipy>=1.3.2 in
c:\users\richard\anaconda3\lib\site-packages (from imbalanced-learn->imblearn)
(1.6.2)
Requirement already satisfied: scikit-learn>=1.0.2 in
c:\users\richard\anaconda3\lib\site-packages (from imbalanced-learn->imblearn)
(1.1.3)
Requirement already satisfied: numpy>=1.17.3 in
c:\users\richard\anaconda3\lib\site-packages (from imbalanced-learn->imblearn)
(1.22.2)
Requirement already satisfied: joblib>=1.1.1 in
c:\users\richard\anaconda3\lib\site-packages (from imbalanced-learn->imblearn)
(1.1.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\richard\anaconda3\lib\site-packages (from imbalanced-learn->imblearn)
(2.2.0)
Installing collected packages: imbalanced-learn, imblearn
Successfully installed imbalanced-learn-0.10.1 imblearn-0.0
Note: you may need to restart the kernel to use updated packages.
```

[48]:
```
# Import the RandomUnderSampler class
from imblearn.under_sampling import RandomUnderSampler

# Split the data into features and labels
X = train_data[['tweet']]
y = train_data['label']

# Create a RandomUnderSampler/RandomOverSampler object
undersampler = RandomUnderSampler()

# transform the data
X_undersampled, y_undersampled = undersampler.fit_resample(X, y)

# Create a downsampled/upsampled dataframe
data_downsampled = pd.DataFrame({'tweet': X_undersampled['tweet'], 'label':␣
 ↪y_undersampled})
```

[49]:
```
def clean_text(text):

```

```
    # tokenize
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(text)


    ## lemmatize + lowercase
    lemmatizer = WordNetLemmatizer()
    for word in text.split():
          token = lemmatizer.lemmatize(word.lower(), pos='v')


    ## remove stopwords
    keep_words = [token for token in tokens if token not in stopwords.
 ↪words('english')]
    row_text = ' '.join(keep_words)
    row_text = ' '.join([word for word in row_text.split() if len(word)>1])  ##␣
 ↪remove one letter words
    row_text = re.sub(r'\w*\d\w*', '', row_text).strip()


    return row_text
```

[50]:
```
train2 = data_downsampled.copy()
train2.head()
```

[50]:

	tweet	label
0	literally geeking so hard over this @user sche…	0
1	our two presidential candidates, everybody. th…	0
2	@user we're seeing @user tonight!! how amazing…	0
3	sex stories ns porn	0
4	saw #indians 2outs away, #browns just miss sb,…	0

[51]:
```
## apply the clean_text function to the 'tweet' column
train2['tweet'] = train2['tweet'].apply(clean_text)
```

[52]:
```
## remove non-ASCII characters
train2.replace(regex=True, to_replace =['¢','€','£', 'Ã', '¬', 'Ð',␣
 ↪'±','½','©','•',
                                        '¾', '§', '¥', '«', '¤', '−', 'œ', '¡', '"',
                                        '|', 'â', ' ', 'Ã', 'Î', '¿', 'µ', '´', '‡',
                                        '»', 'Ž', '®', 'º', 'Ï', 'f', '¶', '¹', ' ',
                                        'á', 'Γ', 'Ç', 'Ö', 'ð'], value='',␣
 ↪inplace=True)
```

[53]:
```
## training-testing data
X_train, X_test, y_train, y_test = train_test_split(train2['tweet'],␣
↪train2['label'], test_size=0.2)
```

[54]:
```
## convert to pandas dataframes
X_train = pd.DataFrame(X_train)
y_train = pd.DataFrame(y_train)

X_test = pd.DataFrame(X_test)
y_test = pd.DataFrame(y_test)
```

Feature Extraction using TF-IDF, short for term frequency–inverse document frequency. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

[56]:
```
# Vectorize the training data using the TfidfVectorizer
vectorizer = TfidfVectorizer(max_features= 1000)
X_train2 = vectorizer.fit_transform(X_train['tweet']).toarray()
X_test2 = vectorizer.fit_transform(X_test['tweet']).toarray()
```

[57]:
```
# Convert the dense arrays to a tensors
train_input_tensor = torch.tensor(X_train2)
test_input_tensor = torch.tensor(X_test2)
```

3.0.1 Simple Transformer Model

[58]:
```
class Transformer(nn.Module):
    def __init__(self, input_size, output_size, num_layers, num_heads,␣
↪hidden_size):
        super(Transformer, self).__init__()
        self.input_size = input_size
        self.output_size = output_size
        self.num_layers = num_layers
        self.num_heads = num_heads
        self.hidden_size = hidden_size

        self.encoder = nn.TransformerEncoder(
            nn.TransformerEncoderLayer(d_model=input_size, nhead=num_heads,␣
↪dim_feedforward=hidden_size),
            num_layers=num_layers,
        )
        self.output_layer = nn.Linear(input_size, output_size)

    def forward(self, input_tensor):

        # Pass the input tensor through the encoder
        encoded_output = self.encoder(input_tensor)
```

```
        # Pass the encoded output through the output layer
        logits = self.output_layer(encoded_output)

        return logits
```

[59]:
```
# Define the model
model = Transformer(input_size=1000, output_size=2, num_layers=2, num_heads=8,␣
 ↪hidden_size=128)
```

[60]:
```
# Convert the labels to a tensor
train_labels = torch.tensor(y_train['label'].values)

# Combine the input tensors into a single dataset
train_dataset = TensorDataset(train_input_tensor, train_labels)

# Create a dataloader from the dataset
train_dataloader = DataLoader(train_dataset, batch_size=8, shuffle=True)
```

[61]:
```
# Convert the labels to a tensor
test_labels = torch.tensor(y_test['label'].values)

# Combine the input tensors into a single dataset
test_dataset = TensorDataset(test_input_tensor, test_labels)

# Create a dataloader from the dataset
test_dataloader = DataLoader(test_dataset, batch_size=1, shuffle=True)
```

[62]:
```
# Define the loss function and the optimizer
loss_fn = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(model.parameters(), lr=0.001)
```

[63]:
```
## Training

num_epochs = 5
model.train()
# Define the training loop
for epoch in range(num_epochs):
  for input_tensor, labels in train_dataloader:
    # Zero the gradients
    optimizer.zero_grad()

    # Forward pass
    logits = model(input_tensor.type(torch.float32))
    loss = loss_fn(logits, labels)

    # Backward pass
```

```
    loss.backward()
    optimizer.step()

  print(f'Epoch {epoch+1}/{num_epochs}, Loss: {loss.item()}')
```

```
Epoch 1/5, Loss: 0.7618595361709595
Epoch 2/5, Loss: 0.6026470065116882
Epoch 3/5, Loss: 0.7132255434989929
Epoch 4/5, Loss: 0.4668183922767639
Epoch 5/5, Loss: 0.6889235973358154
```

[64]:
```
## training and evaluation dataframes
train_df = pd.DataFrame({'tweet': X_train['tweet'], 'label': y_train['label']})
eval_df = pd.DataFrame({'tweet': X_test['tweet'], 'label': y_test['label']})
```

[ ]:
```
# Optional model configuration
model_args = ClassificationArgs(num_train_epochs=1)

# Create a ClassificationModel
model = ClassificationModel(
    "roberta", "roberta-base", args=model_args, use_cuda=False)

# Train the model
model.train_model(train_df,overwrite_output_dir=True)
```

4 Model Inference

[43]:
```
# Test for Hate Speech using predetermined tweets
def predict(inp):
    inp = pd.Series(inp)
    yhat = ((np.ravel(model.predict(inp)).tolist()))
    if yhat[-1] == 1:
        print('Conc : Hate Speech Detected!')
    if yhat[-1] == 0:
        print('Conc: No Hate Speech Detected')


inp = [' Minorities are bad! Fuck off! ', 'Rugby is a fun game! ']
for i in inp:
    print('Entered_Comment: ', i)
    predict(i)
    print('--'*30)
```

```
Entered_Comment:   Minorities are bad! Fuck off!
Conc : Hate Speech Detected!
------------------------------------------------------------
Entered_Comment:  Rugby is a fun game!
```

```
Conc: No Hate Speech Detected
------------------------------------------------------------
```

5 Model Deployment

[67]:
```python
# import required libraries
import tweepy
import time
import pandas as pd
pd.set_option('display.max_colwidth', 1000)

# api key
api_key = "9D6LvvCirf5d16SudvkS2SiKf"
# api secret key
api_secret_key = "5RIuJlxMfsc3drxlwibWc5qgyf2rZPPb9ZPNcTsBFDKkTEx0gf"
# access token
access_token = "1577475918192201730-nzndXCXcumuxdZfj2DkGZTrXnIxGYT"
# access token secret
access_token_secret = "bzqvlVwjcKF4h13Rwcu35aX5JvxGHVUQQ1Z65TmdNDk2G"

# authorize the API Key
authentication = tweepy.OAuthHandler(api_key, api_secret_key)

# authorization to user's access token and access token secret
authentication.set_access_token(access_token, access_token_secret)

# call the api
api = tweepy.API(authentication, wait_on_rate_limit=True)
```

[68]:
```python
def get_related_tweets(text_query):
    # list to store tweets
    tweets_list = []
    # no of tweets
    count = 50
    try:
        # Pulling individual tweets from query
        for tweet in api.search_tweets(q=text_query, count=count):
            print(tweet.text)
            # Adding to list that contains all tweets
            tweets_list.append({'created_at': tweet.created_at,
                                'tweet_id': tweet.id,
                                'tweet_text': tweet.text})
        return pd.DataFrame.from_dict(tweets_list)

    except BaseException as e:
        print('failed on_status,', str(e))
        time.sleep(3)
```

[71]:
```
from get_tweets import get_related_tweets
```

[72]:
```
# importing the required libraries
from flask import Flask, render_template, request, redirect, url_for
from joblib import load


# load the pipeline object
pipeline = load("text_classification.joblib")

# function to get results for a particular text query
def requestResults(name):
    # get the tweets text
    tweets = get_related_tweets(name)
    # get the prediction
    tweets['prediction'] = pipeline.predict(tweets['tweet_text'])
    # get the value counts of different labels predicted
    data = str(tweets.prediction.value_counts()) + '\n\n'
    return data + str(tweets)
```

[73]:
```
# start flask
app = Flask(__name__)

# render default webpage
@app.route('/')
def home():
    return render_template('home.html')

# when the post method detect, then redirect to success function
@app.route('/', methods=['POST', 'GET'])
def get_data():
    if request.method == 'POST':
        user = request.form['search']
        return redirect(url_for('success', name=user))

# get the data for the requested query
@app.route('/success/<name>')
def success(name):
    return "<xmp>" + str(requestResults(name)) + " </xmp> "
```

[ ]:
```
app.run(debug=False)
```

```
 * Serving Flask app '__main__' (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production
deployment.
   Use a production WSGI server instead.
```

```
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [25/Jan/2023 12:28:48] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [25/Jan/2023 12:28:49] "GET /favicon.ico HTTP/1.1" 404 -
127.0.0.1 - - [25/Jan/2023 12:29:00] "POST / HTTP/1.1" 302 -
127.0.0.1 - - [25/Jan/2023 12:29:02] "GET /success/Trump HTTP/1.1" 200 -
```

@BlackerbyNeal @ericareport 1. It was absolutely an insurrection. Members of the Oathkeepers have already been conv… https://t.co/H2Iw5iVucN
RT @RonFilipkowski: Ronny's first tweet about fentanyl ever was on July 3, 2021. He has made 18 tweets since that time blaming the border f…
@The_Trump_Train Hell Yeah!!
@KariLake Are you trying to brake Trump's record of losing?
RT @BlackKnight10k: Now that they've arrested the FBI agent who covered for Trump in 2016 and framed Hillary right before the election, we…
@The_Trump_Train Yes I do.
RT @BlackKnight10k: Now that they've arrested the FBI agent who covered for Trump in 2016 and framed Hillary right before the election, we…
God, I hate #KevinMcCarthy

Please God, use all your power to smite him so we can rid our country of the stench of… https://t.co/HDu7rARP0k
RT @AntonioSabatoJr: Remember Trump had declassification authority. Biden and Pence didn't.
@POTUS Funny how Trump did that in 2020.. lol
RT @MarshaBlackburn: The FBI agent who led the phony investigation into President Trump and Russia has been arrested for his own ties to th…
RT @realmichaelseif: Donald Trump, Joe Biden, Mike Pence all took documents from the White House and the only one who woke up one morning t…
RT @OccupyDemocrats: BREAKING: Fulton County DA announces that the decision whether or not to indict Donald Trump for trying to illegally o…
RT @lindyli: Instead of telling us 7 yrs ago that he was being pressured to sabotage Hillary Clinton, James Comey wrote a book instead

Ins…
@DogRightGirl But the lame stream media tries to make it all the same except they say Trump didn't cooperate
@harryjsisson Yeah sounds like you're starting to wake up. This is good. Just watch and listen closely to those who oppose Trump.
@The_Trump_Train Or "in" his back door
RT @ewarren: If Republicans hadn't spent nearly $2 trillion on the Trump tax cuts, and if they hadn't made it easier for rich people to che…
RT @RonFilipkowski: Ronny hasn't been this excited since Trump stripped down for his annual physical. https://t.co/150dN2vHGM
RT @mrddmia: Trump (lawfully) with his classified presidential records: "Jail the traitor!"

Biden with (stolen) classified records: "Every…

RT @DoliaEstevez: La posibilidad de un nuevo ataque terrorista estilo 9/11 vendría de México, dice Mike Pompeo, ex director de la CIA y ex…
RT @DashDobrofsky: There it is. Classified documents were found at Mike Pence's home in Indiana. This cancels out Joe Biden's classified do…
@LoriMills4CA42 @SpeakerMcCarthy @HouseIntel https://t.co/w1wPHWckfX
@Breaking911 Didn't Trump want Pence hung lol
RT @WalshFreedom: Do I believe Biden knew he had classified documents in his home? No.

Do I believe Pence knew he had classified documents…
Biden, Pelosi, Schumer Suffer Major Defeat #MAGA #TRUMP2024 #Trump

https://t.co/olwsTyvQxK
@Brink_Thinker Trump is back
@laurenboebert A Consensual Search means you agree they have a right to do so. A "Trump is a Crybaby" Search means… https://t.co/W5x0uNm6oF
RT @MarkSZaidEsq: Told ya.

Now former VP Pence should allow @FBI to voluntarily search his offices & residence for additional classified d…
@The_Trump_Train Yes
@GOP Decontrolling gasoline prices. Yeah real smart fucking move. Look it up 1981. You can shove that up your Trump hole
RT @mitchellvii: Allow me to play devil's advocate on the vaccine:

Imagine for a moment that President Trump has refused development of th…
RT @DineshDSouza: BREAKING: FBI official who investigated Trump ties to Russia was just arrested for illegal ties to Russian oligarch
https…
RT @RepAdamSchiff: Kevin McCarthy just kicked me and @RepSwalwell off the Intelligence Committee.

This is petty, political payback for inv…
RT @RepAdamSchiff: Kevin McCarthy just kicked me and @RepSwalwell off the Intelligence Committee.

This is petty, political payback for inv…
RT @MI_James57: Why is raising the debt limit suddenly a huge problem for the GOP?

They did it three times under Trump without objection.
Georgia prosecutor says the decision to charge Trump in 2020 election case is 'imminent' as she urges judge to keep grand jury report secret
Donald Trump has announced that Mike Pence must hang for taking classified documents. He is one persistent former p… https://t.co/MLgJ6fkEEm
@NewYorkStateAG When u ran for office u said u will prosecute Trump that's was ur platform before u know what the c… https://t.co/N5I03FdpBf
RT @l78lancer: Anyone who has been aware of and followed Fani Willis over the

```
past ten years knows that RICO prosecution is her home court.…
RT @blueheartedly: Who thinks Trump gave classified information to Vladimir
Putin?
@realMAGAMAFIA @KariLake @LadyladyHeidiAB yes, it's always all these other ppl
holding you all back and not you all… https://t.co/Zn9SBUtphG
@The_Trump_Train Yes
RT @cathyob1: @SpeakerMcCarthy @HouseIntel You have no morals, no scruples and
no decency. You are not worthy to shine Adam Schiff's shoes.…
RT @IndependentWat6: @DC_Draino @RepAdamSchiff @RepSwalwell Trump is the traitor
to this country along with the sycophant who helped with t…

127.0.0.1 - - [25/Jan/2023 12:29:02] "GET /favicon.ico HTTP/1.1" 404 -
```

5.1 Performance Evaluation & Reporting

In the Twitter Sentiment Analysis project, our group decided to each test and evaluate a Natural Language Processing model. Richard chose to use XGBoost and SKLearn to model the data and evaluate for hate speech with the results being evaluated by Receiever Operating Characteristic or (ROC) curve is a graphical representation of the performance of a binary classification model.

Chris has chosen to use a simple transformer model using the NLTK, SKLearn, and SimpleTransformer libraries. The evaluation of this model uses the Accuracy metric, which is a simple metric that measures the proportion of predictions that are correct. It is commonly used for classification tasks such as text classification and named entity recognition.

5.1.1 XGBoost and SKLearn Model Description

XGBoost (eXtreme Gradient Boosting) is an open-source software library that provides an efficient implementation of the gradient boosting algorithm, which is used for supervised learning tasks such as classification and regression.

The gradient boosting algorithm is an ensemble technique that combines multiple weak models, such as decision trees, to form a strong predictive model. XGBoost specifically uses decision tree ensembles and improves upon the traditional gradient boosting algorithm by implementing several performance enhancements such as regularization and parallel computing.

XGBoost is known for its high performance and scalability, and it has been used in many winning solutions in machine learning competitions. It is also widely used in industry for tasks such as customer segmentation, anomaly detection, and click prediction. It also have various parameter which can be tune to improve the performance.

XGBoost is not typically used in natural language processing (NLP) tasks directly, as it is primarily a gradient boosting algorithm for supervised learning tasks such as classification and regression. However, it can be used as a component in a larger NLP pipeline by training a model using XGBoost and then using the trained model for prediction on NLP tasks.

For example, XGBoost can be used to train a model for text classification tasks such as sentiment analysis and spam detection. The input features for the model can be the bag-of-words representation of the text, and the output label can be the sentiment (positive, negative, neutral) or the class (spam or not spam).

XGBoost can also be used for sequence labeling tasks such as named entity recognition (NER). In this case, the input would be a sequence of words, and the output label would be the entity type (person, location, organization, etc.) for each word in the sequence.

XGBoost's high performance and scalability make it a good choice for NLP tasks that have a large amount of training data, and it is also useful for tasks where interpretability of the model is important.

5.1.2 XGBoost Evaluation

XGBoost is evaluated by ROC or Reciever Operating Characteristic. In statistics, a Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It is a plot of the true positive rate (TPR) against the false positive rate (FPR) for different threshold settings of the model.

The true positive rate, also known as sensitivity or recall, is the proportion of actual positive cases that are correctly classified by the model. The false positive rate, also known as the fall-out, is the proportion of actual negative cases that are incorrectly classified as positive by the model.

An ROC curve is a useful tool for evaluating a model's performance because it shows the trade-off between the true positive rate and the false positive rate. A model with a high true positive rate and a low false positive rate will have a high overall accuracy and will be located in the top left corner of the ROC curve.

A perfect model will have a ROC curve that is a step function with the false positive rate going from 0 to 1 and the true positive rate going from 1 to 1. A random model will have a ROC curve that is diagonal line, for this type of model the true positive and false positive rate are the same.

The area under the ROC curve (AUC) is a measure of the overall performance of the model. A model with an AUC of 1.0 is a perfect model, while a model with an AUC of 0.5 is no better than random.

5.1.3 Simple Transformer Model Description

A transformer model is a type of neural network architecture primarily used for natural language processing tasks such as language translation and text generation. It was introduced in the paper "Attention Is All You Need" by Google researchers in 2017.

A simple transformer model consists of an encoder and a decoder. The encoder takes in a sequence of input tokens (such as words in a sentence) and produces a set of hidden states, which summarize the information in the input sequence. The decoder then takes in these hidden states and generates an output sequence (such as a translation of the input sentence).

Both the encoder and decoder are made up of multiple layers of a type of attention mechanism called self-attention, as well as feed-forward neural network layers. The self-attention mechanism allows the model to weigh the importance of different parts of the input and output sequences, which is useful for tasks such as language translation where certain words or phrases in the source sentence may be more important to the meaning of the target sentence.

5.1.4 Simple Transformer Model Evaluation

Evaluating a transformer model in natural language processing (NLP) tasks involves using metrics that are specific to the task at hand. A common metrics used to evaluate transformer models in NLP tasks includes "Accuracy". Accuracy is a simple metric that measures the proportion of predictions that are correct. It is commonly used for classification tasks such as text classification and named entity recognition.

It's also important to note that in some NLP tasks, it's important to evaluate the model with human evaluation as well, where a group of human evaluators rate the output of the model. In addition to these metrics, it is also important to evaluate the model's performance on a variety of test sets to ensure that it generalizes well to unseen data.

6 Sentiment Analysis Results

6.0.1 XGBoost Model

Our XGBoost model has shown itself to be a prime candidate for sentiment analysis in NLP. The results are shown two-fold, our graphical representation demonstrates the ROC results show a step function with false positive rate starting at 0 and going to 1 which represents a perfect model. Furthermore, our ROC curve shows the true positive and false positive rates are the same as demonstrated by the diagonal line in our graph.

We also calculated an AUC-score of 0.9733875328219913 and F1-score of 0.9159862338889263. The closer our AUC score is to 1 the more perfect the model. The F1 score is a measure of a model's precision and recall. It is commonly used for tasks where the class distribution is imbalanced and accuracy may not be a good indicator of model performance.

Overall, we feel the XGBoost model using SKLearn libraries is a great NLP model for sentiment analysis which has demonstrated reliability and precision.

7 Model Deployment - FLASK

Deploying a natural language processing (NLP) model on Flask is a way to make the model accessible via a web API. Flask is a lightweight web framework for Python that allows for easy creation of web applications and APIs.

The following is a high-level overview of the steps we used to deploy an NLP model on Flask:

- Train and save the NLP model: Train the NLP model using the desired dataset and save the model to a file. This can be done using a library such as PyTorch or TensorFlow.
- Create a Flask application: Create a new Flask application and define the routes that the API will handle. These routes will be used to handle incoming requests and return the model's predictions.
- Load the model: Load the trained NLP model into the Flask application. This can be done using the appropriate library's load function, such as torch.load() for PyTorch models.
- Define the prediction endpoint: Define a route that takes in input data, passes it through the loaded model, and returns the model's predictions. This endpoint will handle the request and response for the API.

- Start the server: Start the Flask development server to make the API accessible. This can be done using the app.run() method.
- Test the API: Test the API by making requests to the endpoint and checking the returned predictions.

It's also possible to deploy the model using a production-ready web server such as Gunicorn or uWSGI and use a reverse proxy server like Nginx to handle the request and response on the API.

It's also important to consider security and access control when deploying the model, you should protect the API endpoint with a token-based authentication or API key to prevent unauthorized access.