!pip install lime

```
Requirement already satisfied: lime in /usr/local/lib/python3.7/dist-packages (0.2.0.1)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from lime) (4.62.3)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from lime) (1.4.1)
Requirement already satisfied: scikit-image>=0.12 in /usr/local/lib/python3.7/dist-packages (from lime) (0.18.
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from lime) (3.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from lime) (1.21.5)
Requirement already satisfied: scikit-learn>=0.18 in /usr/local/lib/python3.7/dist-packages (from lime) (1.0.2
Requirement already satisfied: imageio>=2.3.0 in /usr/local/lib/python3.7/dist-packages (from scikit-image>=0.
12->lime) (2.4.1)
Requirement already satisfied: pillow!=7.1.0,!=7.1.1,>=4.3.0 in /usr/local/lib/python3.7/dist-packages (from s
cikit-image>=0.12->lime) (7.1.2)
Requirement already satisfied: PyWavelets>=1.1.1 in /usr/local/lib/python3.7/dist-packages (from scikit-image>
=0.12->lime) (1.2.0)
Requirement already satisfied: tifffile>=2019.7.26 in /usr/local/lib/python3.7/dist-packages (from scikit-imag
e>=0.12->lime) (2021.11.2)
Requirement already satisfied: networkx>=2.0 in /usr/local/lib/python3.7/dist-packages (from scikit-image>=0.1
2->lime) (2.6.3)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packa
ges (from matplotlib->lime) (3.0.7)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib
->lime) (2.8.2)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->lime)
(0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->1
ime) (1.3.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.1->
matplotlib->lime) (1.15.0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit-learn>=0.18
->lime) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from scikit-lea
rn >= 0.18 - lime) (3.1.0)
                                                                                                           In []:
```

!pip install tensorflow_text

```
Requirement already satisfied: tensorflow_text in /usr/local/lib/python3.7/dist-packages (2.8.1)
Requirement already satisfied: tensorflow<2.9,>=2.8.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow_text) (2.8.0)
Requirement already satisfied: tensorflow-hub>=0.8.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow)
```

 w_{text}) (0.12.0) Requirement already satisfied: typing-extensions>=3.6.6 in /usr/local/lib/python3.7/dist-packages (from tensor

flow<2.9,>=2.8.0->tensorflow_text) (3.10.0.2)
Requirement already satisfied: wrapt>=1.11.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=

2.8.0->tensorflow_text) (1.13.3)

Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in /usr/local/lib/python3.7/dist-packages

(from tensorflow<2.9,>=2.8.0->tensorflow_text) (0.24.0) Requirement already satisfied: opt-einsum>=2.3.2 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.8.0->tensorflow text) (3.3.0)

Requirement already satisfied: grpcio<2.0,>=1.24.3 in /usr/local/lib/python3.7/dist-packages (from tensorflow< 2.9,>=2.8.0->tensorflow text) (1.43.0)

Requirement already satisfied: termcolor>=1.1.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.8.0->tensorflow text) (1.1.0)

Requirement already statis (id: tensorboard<2.9,>=2.8 in /usr/local/lib/python3.7/dist-packages (from tensorflo

w<2.9,>=2.8.0->tensorflow_text) (2.8.0)
Requirement already satisfied: tf-estimator-nightly==2.8.0.dev2021122109 in /usr/local/lib/python3.7/dist-pack

ages (from tensorflow<2.9,>=2.8.0->tensorflow_text) (2.8.0.dev2021122109)

Requirement already satisfied: protobuf>=3.9.2 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,

>=2.8.0->tensorflow_text) (3.17.3)

Requirement already satisfied: gast>=0.2.1 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.8.0->tensorflow_text) (0.5.3)

Requirement already satisfied: google-pasta>=0.1.1 in /usr/local/lib/python3.7/dist-packages (from tensorflow< 2.9, >=2.8.0->tensorflow_text) (0.2.0)

Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.8.0->tensorflow text) (1.21.5)

Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.8 .0->tensorflow text) (57.4.0)

Requirement already satisfied: h5py>=2.9.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.

```
8.0->tensorflow text) (3.1.0)
Requirement already satisfied: astunparse>=1.6.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.
9,>=2.8.0->tensorflow text) (1.6.3)
Requirement already satisfied: keras<2.9,>=2.8.0rc0 in /usr/local/lib/python3.7/dist-packages (from tensorflow
<2.9,>=2.8.0->tensorflow text) (2.8.0)
Requirement already satisfied: six>=1.12.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>=2.
8.0->tensorflow text) (1.15.0)
Requirement already satisfied: keras-preprocessing>=1.1.1 in /usr/local/lib/python3.7/dist-packages (from tens
orflow\langle 2.9, \rangle = 2.8.0 - \text{tensorflow text} (1.1.2)
Requirement already satisfied: absl-py>=0.4.0 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,>
=2.8.0->tensorflow text) (1.0.0)
Requirement already satisfied: libclang>=9.0.1 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.9,
>=2.8.0->tensorflow_text) (13.0.0)
Requirement already satisfied: flatbuffers>=1.12 in /usr/local/lib/python3.7/dist-packages (from tensorflow<2.
9, \geq 2.8.0 - \text{tensorflow text} (2.0)
Requirement already satisfied: wheel<1.0,>=0.23.0 in /usr/local/lib/python3.7/dist-packages (from astunparse>=
1.6.0->tensorflow<2.9,>=2.8.0->tensorflow text) (0.37.1)
nsorflow < 2.9, >= 2.8.0 - tensorflow text) (1.5.2)
Requirement already satisfied: werkzeug>=0.11.15 in /usr/local/lib/python3.7/dist-packages (from tensorboard<2
.9,>=2.8-tensorflow<2.9,>=2.8.0->tensorflow text) (1.0.1)
Requirement already satisfied: tensorboard-data-server<0.7.0,>=0.6.0 in /usr/local/lib/python3.7/dist-packages
(from tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (0.6.1)
Requirement already satisfied: google-auth<3,>=1.6.3 in /usr/local/lib/python3.7/dist-packages (from tensorboa
rd<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (1.35.0)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in /usr/local/lib/python3.7/dist-packages (from t
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.7/dist-packages (from tensorboard<2.9
,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow_text) (3.3.6)
Requirement already satisfied: requests<3,>=2.21.0 in /usr/local/lib/python3.7/dist-packages (from tensorboard
<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (2.23.0)
\label{lem:condition} \textbf{Requirement already satisfied: google-auth-oauthlib} < 0.5, >= 0.4.1 \ \text{in /usr/local/lib/python3.7/dist-packages (from the condition of the conditio
m tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (0.4.6)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /usr/local/lib/python3.7/dist-packages (from google-au
\label{thm:continuous} th<3,>=1.6.3-> tensorboard<2.9,>=2.8-> tensorflow<2.9,>=2.8.0-> tensorflow_text) \\ (0.2.8)
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.7/dist-packages (from google-auth<3,>=1
.6.3 - \texttt{>} tensorboard < 2.9, >= 2.8 - \texttt{>} tensorflow < 2.9, >= 2.8.0 - \texttt{>} tensorflow text) \tag{4.8}
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from google-a
uth<3,>=1.6.3->tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (4.2.4)
Requirement already satisfied: requests-oauthlib>=0.7.0 in /usr/local/lib/python3.7/dist-packages (from google
-auth-oauthlib < 0.5, >= 0.4.1 - \\ \\ + 1- \\ \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - \\ + 2.8 - 
Requirement already satisfied: importlib-metadata>=4.4 in /usr/local/lib/python3.7/dist-packages (from markdow
\label{eq:n>=2.6.8-} $$n>=2.6.8-$$ tensorboard<2.9,>=2.8-$$ tensorflow<2.9,>=2.8.0-$$ tensorflow_text) $$ (4.11.0) $$
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metadata>=4
.4->markdown>=2.6.8->tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (3.7.0)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in /usr/local/lib/python3.7/dist-packages (from pyasn1-mod
ules \ge 0.2.1 - google-auth < 3, \ge 1.6.3 - tensorboard < 2.9, \ge 2.8 - tensorflow < 2.9, \ge 2.8.0 - tensorflow text) \quad (0.4.8)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests<3,>
=2.21.0->tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (2021.10.8)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packag
es (from requests<3,>=2.21.0->tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests<3,>=
2.21.0->tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests<3,>=2.21.
0->tensorboard<2.9,>=2.8->tensorflow<2.9,>=2.8.0->tensorflow text) (2.10)
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.7/dist-packages (from requests-oauthl
ib >= 0.7.0 - youngle-auth-oauthlib < 0.5, >= 0.4.1 - yensorboard < 2.9, >= 2.8 - yensorflow < 2.9, >= 2.8.0 - yensorflow text) (
3.2.0)
                                                                                                                                                                                                 In [ ]:
import os
import pandas as pd
```

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
from time import time
warnings.filterwarnings("ignore")
%matplotlib inline
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import re
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
```

```
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad sequences
from keras.models import Sequential
from keras.layers import Dense, Flatten, Embedding, Input, Dropout, LSTM
from keras.utils.np_utils import to_categorical
from tensorflow.python.keras.callbacks import TensorBoard
from sklearn.model selection import train test split
import tensorflow as tf
import tensorflow hub as hub
import tensorflow text as text
import pandas as pd
from sklearn.metrics import f1 score, confusion matrix
from keras.callbacks import EarlyStopping, ReduceLROnPlateau, ModelCheckpoint
from lime.lime_text import LimeTextExplainer
from keras.models import load model
[nltk data] Downloading package stopwords to /root/nltk data...
[nltk data]
               Package stopwords is already up-to-date!
[nltk data] Downloading package punkt to /root/nltk data...
               Package punkt is already up-to-date!
[nltk data]
[nltk data] Downloading package wordnet to /root/nltk data...
[nltk data] Package wordnet is already up-to-date!
                                                                                                                       In [ ]:
from google.colab import drive
drive.mount('/content/drive')
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", fo
rce remount=True).
                                                                                                                       In []:
%cd /content/drive/MyDrive/kaggle toxic/
/content/drive/MyDrive/kaggle toxic
                                                                                                                       In []:
all data = pd.read csv('all data.csv')
                                                                                                                       In []:
all data.head(5)
                                                                                                                      Out[]:
                                 created\_date \quad publication\_id \quad parent\_id \quad article\_id
       id comment text split
                                                                             rating funny wow sad likes disagree toxicity sev
               He got his
                                   2017-03-06
            monev... now
  1083994
                        train
                                                      21
                                                             NaN
                                                                    317120 approved
                                                                                        0
                                                                                             0
                                                                                                 0
                                                                                                               0 0.373134
                             15:21:53.675241+00
            he lies in wait
                  till a...
            Mad dog will
                                   2016-12-02
            surely put the
                                                                                                               0 0.605263
    650904
                        train
                                                             NaN
                                                                     154086 approved
               liberals in
                             16:44:21.329535+00
               mental...
             And Trump
            continues his
                                   2017-09-05
2 5902188
                        train
                                                      55
                                                             NaN
                                                                    374342 approved
                                                                                             0
                                                                                                 2
                                                                                                               7 0 666667
                lifelong
                             19:05:32.341360+00
           cowardice by ...
           "while arresting
                                   2016-11-01
               a man for
3 7084460
                        test
                                                      13
                                                             NaN
                                                                     149218 approved
                                                                                        0
                                                                                             0
                                                                                                 0
                                                                                                       0
                                                                                                               0 0.815789
                resisting
                             16:53:33.561631+00
               arrest".\...
              Tucker and
```

In []:

NaN

344096 approved

0

0 0.550000

toxic = []

5410943

#making comments which have probability more than 0.5 as toxic and marking them as 1 while non-toxic as 0
for i in all_data['toxicity']:

if i > 0.5:

toxic.append(1)

Paul are both

total bad ass

mofo's.

train

2017-06-14

05:08:21.997315+00

```
else:
        toxic.append(0)
all_data['toxic_binary'] = toxic
                                                                                                            In []:
all_data['sub_toxic'] = all_data[['severe_toxicity','obscene','sexual_explicit', 'identity_attack','insult',
                                                                                                            In []:
sub toxic = []
for j in range(len(all data)):
   if all_data['toxic_binary'][j] == 1:
      if all_data['sub_toxic'][j] == 'severe_toxicity':
          sub_toxic.append(6)
      if all data['sub toxic'][j] == 'obscene':
          sub toxic.append(5)
      if all_data['sub_toxic'][j] == 'sexual_explicit':
          sub toxic.append(4)
      if all_data['sub_toxic'][j] == 'identity_attack':
          sub toxic.append(3)
      if all data['sub toxic'][j] == 'insult':
          sub toxic.append(2)
      if all_data['sub_toxic'][j] == 'threat':
          sub_toxic.append(1)
   if all data['toxic binary'][j] == 0:
       sub_toxic.append(0)
all data['sub toxic'] = sub toxic
                                                                                                            In []:
stop = set(stopwords.words('english'))
def clean(text):
  text token = word tokenize(text)
  filtered_text = ''.join([w.lower() for w in text_token if w.lower() not in stop and len(w) > 2])
  filtered_text = filtered_text.replace(r"[^a-zA-Z]+", '')
  text_only = re.sub(r'\b\d+\b', '', filtered_text)
  \verb|clean_text| = text_only.replace(',', '').replace('.', '').replace(':', '')|
  return clean text
                                                                                                            In []:
all data['clean comment'] = [clean(str(x)) for x in all data['comment text']]
Splitting Data
                                                                                                            In []:
train = all data.loc[all data['split']=='train']
```

train.head(5)

	id	comment_text	split	created_date	publication_id	parent_id	article_id	rating	funny	wow	sad	likes	disagree	toxicity	sev
0	1083994	He got his money now he lies in wait till a	train	2017-03-06 15:21:53.675241+00	21	NaN	317120	approved	0	0	0	2	0	0.373134	
1	650904	Mad dog will surely put the liberals in mental	train	2016-12-02 16:44:21.329535+00	21	NaN	154086	approved	0	0	1	2	0	0.605263	
2	5902188	And Trump continues his lifelong cowardice by	train	2017-09-05 19:05:32.341360+00	55	NaN	374342	approved	1	0	2	3	7	0.666667	
4	5410943	Tucker and Paul are both total bad ass mofo's.	train	2017-06-14 05:08:21.997315+00	21	NaN	344096	approved	0	0	0	1	0	0.550000	
5	6290444	Cry me a river, why don't you.\nDrinking, drug	train	2017-11-04 22:04:11.596185+00	54	6290143.0	396946	rejected	0	0	0	0	0	0.203390	
4														lr	• n []:

test = all_data.loc[all_data['split']=='test']
test.head(5)

Out[]:

															F 1.
	id	$comment_text$	split	created_date	publication_id	parent_id	article_id	rating	funny	wow	sad	likes	disagree	toxicity	se
3	7084460	"while arresting a man for resisting arrest".\	test	2016-11-01 16:53:33.561631+00	13	NaN	149218	approved	0	0	0	0	0	0.815789	
10	7141509	NO! There are no alternative facts. Go check	test	2017-01-30 02:53:48.012277+00	21	919529.0	164687	approved	1	0	0	0	0	0.597222	
11	7077814	the more you whine sore loser Artster\n\nthe m	test	2016-12-03 00:17:42.300700+00	54	649753.0	154126	approved	0	0	0	0	0	0.650000	
38	7147990	There's rarely opportunity to agree with Benne	test	2017-09-13 16:37:16.990602+00	102	NaN	377304	approved	1	0	0	1	2	0.111111	
42	7008066	The Law has every freedom to be an asss!	test	2017-07-09 07:03:44.153492+00	54	5556167.0	353158	approved	0	0	0	0	0	0.800000	
4															F

train = train.reset_index(drop=True)
test = test.reset_index(drop=True)

ln []:

In []:

X = train['clean_comment']
Y = train['sub_toxic']

In []:

x_train, x_test, y_train, y_test = train_test_split(X.values, Y.values, test_size=0.2, stratify=Y)

In []:

x_train.shape, x_test.shape, y_train.shape, y_test.shape

Out[]:

```
train_labels = to_categorical(y_train)
val labels = to categorical(y test)
```

Loading BERT

In []:
bert_preprocess = hub.KerasLayer("https://tfhub.dev/tensorflow/bert_en_uncased_preprocess/3")
bert_encoder = hub.KerasLayer("https://tfhub.dev/tensorflow/small_bert/bert_en_uncased_L-4_H-512_A-8/2") #bert
In []:
bert = load_model('bert.hdf5', custom_objects={'KerasLayer': bert_preprocess})
In []:

bert.summary()

Model: "model"

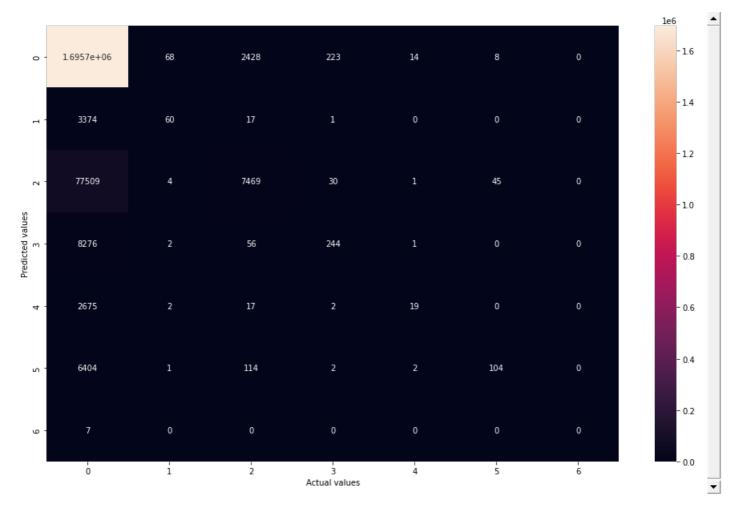
Layer (type)	Output Shape	Param #	Connected to
text (InputLayer)	[(None,)]	0	[]
keras_layer (KerasLayer)	<pre>{'input_type_ids': (None, 128), 'input_word_ids': (None, 128), 'input_mask': (None, 128)}</pre>		['text[0][0]']
keras_layer_1 (KerasLayer)	{'default': (None, 512), 'pooled_output': (None, 512), 'encoder_outputs': [(None, 128, 512), (None, 128, 512), (None, 128, 512), (None, 128, 512)], 'sequence_output': (None, 128, 512)}		['keras_layer[0][0]', 'keras_layer[0][1]', 'keras_layer[0][2]']
dropout (Dropout)	(None, 512)	0	['keras_layer_1[0][5]']
output (Dense)	(None, 7)	3591	['dropout[0][0]']

Total params: 28,767,240 Trainable params: 3,591

Non-trainable params: 28,763,649

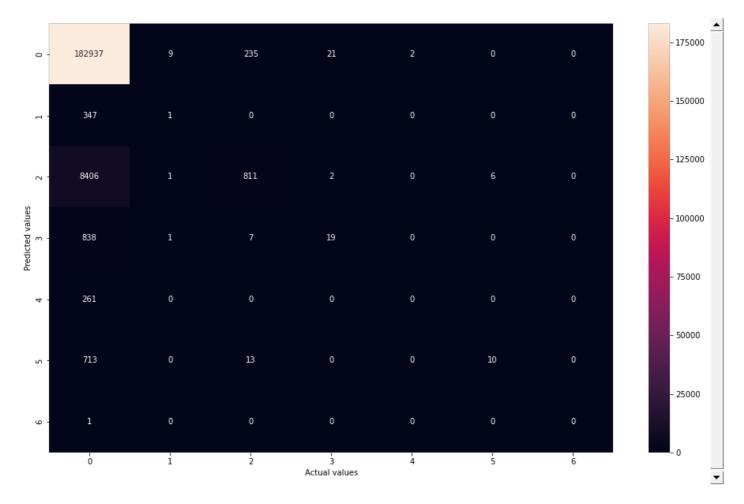
Train Prediction

```
confusion = confusion_matrix(train['sub_toxic'].values, train_classes)
plt.figure(figsize = (16,10))
sns.heatmap(confusion, annot=True, fmt='g')
plt.xlabel('Actual values')
plt.ylabel('Predicted values')
plt.show()
```



Test Prediction

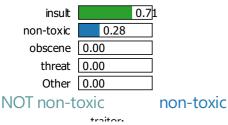
```
In []:
test prediction = bert.predict(test['clean comment'])
                                                                                                            In []:
test_classes = np.argmax(test_prediction,axis=1)
                                                                                                            In []:
fone2 = f1_score(test['sub_toxic'].values, test_classes, average=None)
fone2
                                                                                                           Out[]:
array([0.9712429 , 0.00555556, 0.15759813, 0.04189636, 0.
       0.02659574, 0.
                                                                                                            In []:
confusion = confusion_matrix(test['sub_toxic'].values, test_classes)
plt.figure(figsize = (16,10))
sns.heatmap(confusion, annot=True, fmt='g')
plt.xlabel('Actual values')
plt.ylabel('Predicted values')
plt.show()
```



Train: Incorrect predictions (Worst Case)

LIME analysis : Class 0

```
In []:
train['predicted'] = train classes
                                                                                                             In []:
new_df = train.loc[train['sub_toxic']==0]
new_df = new_df[new_df['predicted'] != 0]
new df.reset index(drop=True, inplace=True)
                                                                                                             In []:
idx= np.random.random integers(0,len(new df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 0
Predicted class: 2
                                                                                                             In []:
class names = ['non-toxic','threat','insult','identity attack','sexual explicit','obscene','severe toxicity']
explainer = LimeTextExplainer(class names = class names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
  Prediction probabilities
```



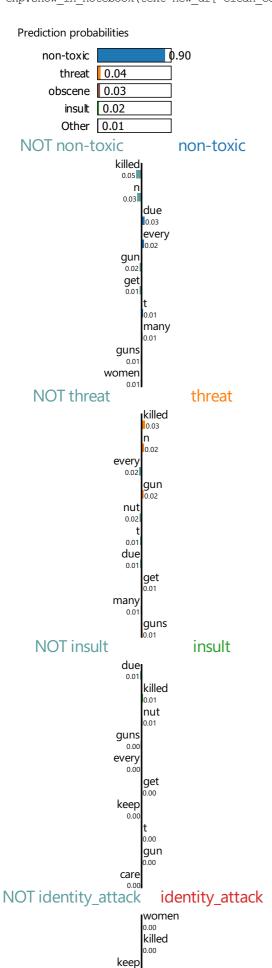
```
11 a11 01 0.19
                   murderous
0.16
                    millionaire
                         0.12
                           like
                         0.10
                      leftwing
                          0.08
                          blow
                          0.07
                        penny
<sub>0.06</sub>
                          send
                               dollars
0.03
      NOT threat
                                          threat
                           fees
0.01
                        dollars
0.01
                    millionaire
                                murderous
                               blow
                        rather
                           0.0
                          legal
                       turned
                            0.00
                        billion
       NOT insult
                                           insult
                               traitor
0.18
                               murderous
0.15
                               millionaire
0.13
                               like
                               0.10
leftwing
                               0.08
blow
                               0.06
penny
0.06
rather
                                0.04
                               turned
dollars
0.03
NOT identity_attack
                                   identity_attack
                    millionaire
                                instead
                       turned
                            0.00
                                leftwing
                                traitor
                           fees
                                murderous
                        penny
                            0.0
                                would
                                0.00
                               send
NOT sexual_explicit
                                 sexual_explicit
                          kadrı
                               blow
                        0.00
```

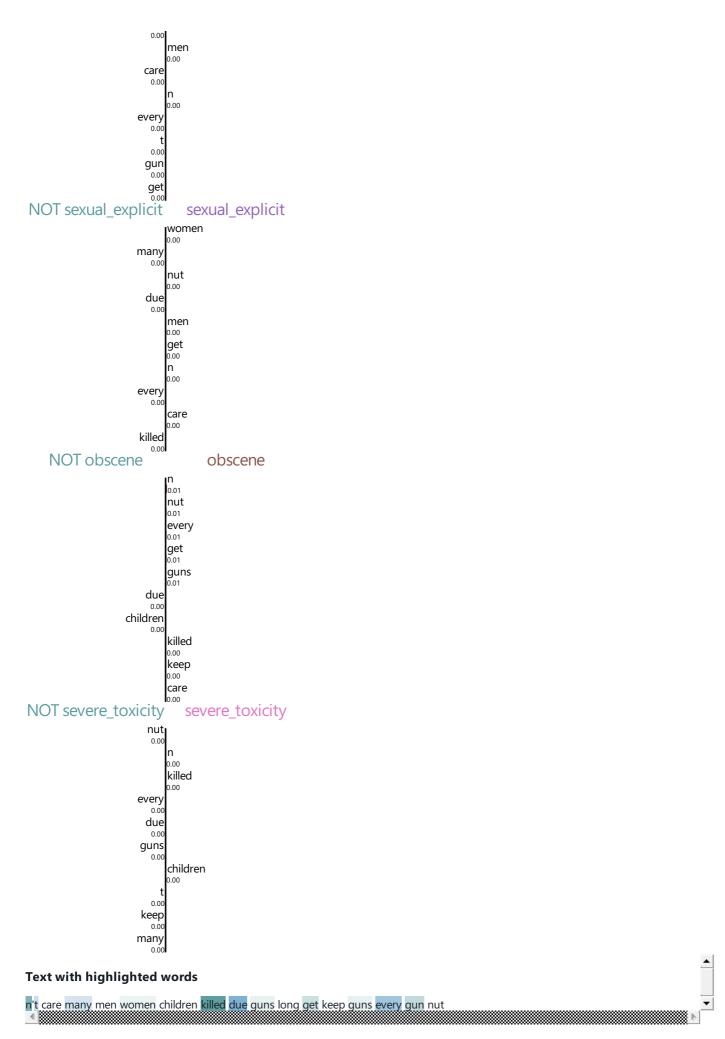
```
noiiiia
                          0.00
                        send
                              legal
                             would
                              0.00
                             murderous
                             millionaire
                      traitor
                        hero
     NOT obscene
                                      obscene
                             ıblow
                     instead
                          0.00
                              turned
                        hero
                        kadı
                          0.0
                      would
                          0.0
                      rather
                          0.0
                             dollars
                             <sub>0.00</sub>
billion
NOT severe_toxicity
                                 severe_toxicity
                             ıkadr
                      dollars
                          0.0
                         fees
                          0.0
                       rather
                  millionaire
                              traitor
                             murderous
                         like
                          0.0
                       would
Text with highlighted words
would rather blow billion dollars legal fees send penny murderous traitor like kadr instead turned leftwing hero millionaire
```

LIME analysis: Class 1

```
In []:
train['predicted'] = train_classes
                                                                                                             In []:
new df = train.loc[train['sub toxic']==1]
new df = new df[new df['predicted'] != 1]
new_df.reset_index(drop=True, inplace=True)
                                                                                                             In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 1
Predicted class: 0
```

class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,exp.show_in_notebook(text=new_df["clean_comment"][idx])





```
train['predicted'] = train classes
                                                                                                                        In []:
new_df = train.loc[train['sub_toxic']==2]
new_df = new_df[new_df['predicted'] != 2]
new_df.reset_index(drop=True, inplace=True)
                                                                                                                        In [ ]:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 2
Predicted class: 0
                                                                                                                        In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class names = class names)
exp = explainer.explain instance(new df["clean comment"][idx], bert.predict, num features = 10,labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
  Prediction probabilities
       non-toxic
                            0.81
           insult 0.10
  identity_attack 0.10
        obscene 0.00
          Other 0.00
   NOT non-toxic
                               non-toxic
                        ezra
                 lunatics
                    right
                    0.07
                     nazi
                         supporters
                         ties
                     hell
                     0.05
                 anyway
                     0.03
                     anti
     NOT threat
                                 threat
                     ezra
                         lnazi
                         0.00
                   media
                      0.0
                         hell
                    right
                  canada
                      0.0
               supporters
                      0.0
                      fai
                      0.0
                         lunatics
     NOT insult
                                  insult
                    ezra
                        lunatics
                        0.07
```

0.06

In []:

```
ties
                 supporters
                        0.05
                            nazi
0.05
                            hell
                             0.04
                            anyway
                     canada
                         far
                         0.02
NOT identity_attack
                               identity_attack
                            nazi
0.02
                            white
                             0.02
                            supremacists
                            0.02
                            ezra
                            anti
                            0.02
                            right
                             0.01
                 supporters
                            0.01
                            groups
                         far
NOT sexual_explicit
                                sexual_explicit
                 supporters
                             0.00
                            lunatics
               supremacists
                       rebel
                      media
                         0.00
                         far
                         0.00
                     canada
                         0.00
                            ezra
                      levant
    NOT obscene
                                    obscene
                 supporters
                            hell
                             0.00
               supremacists
                             right
                        naz
                         0.00
                        anti
                         0.00
                            lunatics
                       wing
                            anyway
                            night
NOT severe_toxicity
                               severe_toxicity
                       right
0.00
                        ezra
0.00
                            supporters
                            0.00
                            media
                            0.00
                            n
<sub>0.00</sub>
```

```
rebel
groups
anyway
oo

it
ooo
far
ooo
far
ooo

Text with highlighted words

rebel media ties right-wing groups n't impossible n't far right lunaties canada right right anyway hell erre levant sleep night knowing well far-right/alt-right white supremacists anti-semites nazi supporters erre

LIME analysis: Class 3

train['predicted'] = train_classes

new_df = train.loc[train['sub_toxic']==3]
new_df = new_df[new_df['predicted'] != 3]
new_df = new_df[new_df['predicted'] != 3]
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
```

In []:

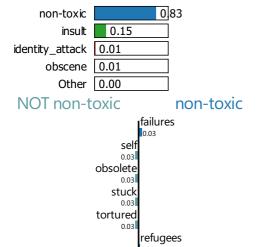
In []:

In []:

explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,exp.show_in_notebook(text=new_df["clean_comment"][idx])

Prediction probabilities

Actual class: 3



NOT threat threat failures

punishment prejudices

respect 0.01 expect 0.01 careers 0.01

```
torturea
                     perceived
                                 dead
                                 enemies
                                 march
                     burdened
                           0.00
21st
        NOT insult
                                             insult
                                obsolete
10.03
self
                                 l<sub>0.02</sub>
stuck
                                 0.02
                                 enemies
                                  0.02
                        respect
                             0.01
                      refugees
                     perceived
                   immigrants
                             0.0
                        careers
                             0.0
expect
0.00

NOT identity_attack
                                     identity_attack
                        failures
0.03
                                 white
                                 boys
                                 0.00
                                 1950s
                                 0.00
                                 male
                                 0.00
                                 nationalist
                                 <sub>0.00</sub>
males
                                 0.00
21st
                                 women
                                tortured
NOT sexual_explicit
                                     sexual_explicit
                        failures
0.00
                       refugees
                                 punishment
                         expect
0.00
                                 males
0.00
                                 women
                                 <sup>0.00</sup>
1950s
                                 0.00
                                 male
                                 0.00
                                 boys
                                 0.00
dress
0.00
     NOT obscene
                                          obscene
                       similarly
                                 punishment
                                 guys
<sub>0.00</sub>
                            age
0.00
                        failures
```

```
perceived
                  prejudices
                            lcruise
                             0.00
                            boys
                            0.00
                            stuck
NOT severe toxicity
                               severe toxicity
                    failures
                    enough
                            lenemies
                            nationalist
                            ideology
                            1950s
                     expect
                    respect
                        0.0
                  whatever
                         0.0
                       boys
```

Text with highlighted words

proud boys describe guys doofusses doofi embittered white males tortured loss privilege lashing perceived enemies refugees 1950s obsolete years age whatever classic fodder far right/white supremacist/alt right/nationalist fringe crushed simply male turned enough put front gravy train stuck dead-end no-hope careers watch women immigrants smart enough hard-working enough educate 21st century cruise past line dress prejudices expect taken seriously anyone similarly self-burdened brethren let march around little uniforms smaller ideology non-existent self-respect punishment failures

LIME analysis : Class 4

photos

```
In []:
train['predicted'] = train classes
                                                                                                              In []:
new df = train.loc[train['sub toxic']==4]
new df = new df[new df['predicted'] != 4]
new_df.reset_index(drop=True, inplace=True)
                                                                                                              In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new df['predicted'][idx])
Actual class: 4
Predicted class:
                                                                                                              In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class names = class names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
  Prediction probabilities
                          0.83
      non-toxic
          insult 0.08
  sexual explicit 0.06
  identity_attack 0.02
         Other 0.01
   NOT non-toxic
                            non-toxic
                 naked
```

```
0.05
                         sex
                         0.04
                   exploiting
                         0.03
                         0.03
                              publishing
                     deleted
0.02
                     women
                         0.02
                       make
      NOT threat
                                       threat
                             ımake
                  publishing
                      naked
                     photos
                   comment
                     opinion
                     fortune
                             could
      NOT insult
                                        insult
                     photos
0.04
                             exploiting
<sub>0.02</sub>
                             deleted
                             guy
                    sensitive
                         0.01
                     women
                             0.01
                             fortune
                             0.01
                             n
<sub>0.01</sub>
NOT identity_attack
                                 identity_attack
                             ınaked
                             0.01
                             women
                             world
                             0.01
                     fortune
                          0.0
                       could
                   comment
                          0.00
                    sensitive
                          0.00
NOT sexual_explicit
                                 sexual_explicit
                             sex
0.04
                             naked
                  publishing
<sub>0.02</sub>
                             women
```

```
world
                           make
                   photos
                       0.0
                   fortune
                       0.01
                   strange
    NOT obscene
                                  obscene
                           deleted
                           world
                           0.00
                          n
                           0.00
                           sex
                           0.00
                          guy
                           comment
                   strange
                           lwomen
                  sensitive
NOT severe_toxicity
                              severe_toxicity
                           photos
                           0.00
                        0.0
                   fortune
                           world
                           0.00
                           women
                           deleted
                           0.00
                           fame
                           0.00
                     make
Text with highlighted words
guy could make fortune fame sex publishing exploiting naked women photos n't make comment practice deleted opinion sensitive naked
photos strange world
```

LIME analysis : Class 5

```
In[]:
train['predicted'] = train_classes

In []:

new_df = train.loc[train['sub_toxic']==5]
    new_df = new_df[new_df['predicted'] != 5]
    new_df.reset_index(drop=True, inplace=True)

In []:

idx= np.random.random_integers(0,len(new_df))
    print('Actual class: ', new_df['sub_toxic'][idx])
    print('Predicted class: ', new_df['predicted'][idx])

Actual class: 5
Predicted class: 0

In []:
```

```
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10, labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
  Prediction probabilities
       non-toxic
                              0.95
           insult 0.03
        obscene 0.01
  identity_attack 0.01
          Other 0.00
   NOT non-toxic
                               non-toxic
                     crap
                     0.03
                         specifically
                    ultra
                      0.02
                          said
                         0.01
                         case
                         might
                         example
                         0.01
                         specific
                         0.01
                   makes
     NOT threat
                                  threat
               specifically
                   makes
                         crap
                          0.00
                     nor
                      0.0
              comparable
                  specific
                         ultra
                         0.00
                  evident
                        paint
0.00
     NOT insult
                                  insult
               specifically
                         racist
                         crap
                     said
0.01
                     case
                      0.01
                         ultra
                 example
                    detai
                      0.01
                         host
                         0.00
                         think
NOT identity_attack
                            identity_attack
                         racist
                         0.00
                         ultra
```

racism _{0.00}

```
unspecified
                             0.00
                             0.00
                             makes
                      detail
                         0.0
                 specifically
                         0.0
                     session
                         0.00
                     specific
NOT sexual_explicit
                                sexual_explicit
                            unspecified
                             racist
                     others
                             specifically
                             supposedly
                             ultra
                             that
                             0.00
                             ln
                             0.00
                             makes
                             0.00
                      sense
    NOT obscene
                                     obscene
                             crap
                             racist
                             0.00
                             ultra
                             0.00
                             makes
                             0.00
                        said
                         0.00
                             non
                             refer
                      might
                        case
                         0.00
                 something
NOT severe toxicity
                                severe toxicity
                            unspecified
                             specifically
                             makes
                             0.00
                            that
                             0.00
                             example
                             0.00
                             n
                             0.00
                             racist
                             define
                             0.00
                        case
                         0.0
                       refer
```

Text with highlighted words

unspecified 'crap refer define specifically makes whatever think 'that something comparable documented racism evident session case reagan romney 'host others supposedly 'non-liberal specific follow politics pretty clearly done several decades enough try paint pictures big brush picture n't make sense unless add detail exactly say 'it said specifically said example might john tepton specifically makes refer 'ultra racist

													Ou	ut[]:
	id	comment_text	split	created_date	$publication_id$	parent_id	article_id	rating	funny	wow	sad	likes	disagree	to
7656	6214744	female assassins for kim	train	2017-10-24 19:47:26.537935+00	55	NaN	392890	approved	0	0	0	0	0	
537774	5801652	Awesome! Lets cut the head off hate! Lets stab	train	2017-08-19 17:33:30.941935+00	21	NaN	368010	rejected	0	0	0	0	0	
1195376	741611	So dealers of death deserve clemency? What a w	train	2016-12-22 07:13:30.077948+00	21	NaN	157754	approved	0	0	0	0	0	
1347411	5640242	I've worked in some estrogen swamps. Let me te	train	2017-07-22 18:36:05.726092+00	54	NaN	357876	rejected	0	0	0	0	0	
1435818	5833628	Let me remind you people: If you disagree with	train	2017-08-24 19:24:09.989776+00	54	NaN	370106	approved	3	0	0	0	0	
1450132	6214411	Peta is busy killing cats and dogs.\nhttps://w	train	2017-10-24 19:08:54.717860+00	21	6212387.0	392564	approved	0	0	0	2	0	
1628400	5963721	Mr. Dries, Bertha Cooper represents the malign	train	2017-09-16 12:21:43.299091+00	85	NaN	378495	rejected	0	0	0	0	0	
4														Þ

LIME analysis : Class 6

```
In []:
train['predicted'] = train_classes

In []:
new_df = train.loc[train['sub_toxic']==6]
new_df = new_df[new_df['predicted'] != 6]
new_df.reset_index(drop=True, inplace=True)

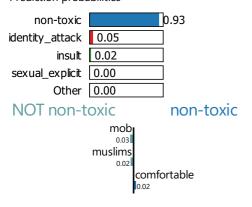
In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])

Actual class: 6
Predicted class: 0

In []:
class_names = ['non-toxic', 'threat', 'insult', 'identity_attack', 'sexual_explicit', 'obscene', 'severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
```

exp = explainer.explain instance(new df["clean comment"][idx], bert.predict, num features = 10,labels=[0, 1,2,

Prediction probabilities



exp.show_in_notebook(text=new_df["clean_comment"][idx])

```
represents
                 malignant
                           designed
                           cooper
                           0.01
                      dries
                        0.01
                    bertha
                        0.01
                     know
                        0.01
     NOT threat
                                    threat
                 represents
                           mob
               comfortable
                        0.0
                        mr
                        0.00
                  anything
                        0.0
                     never
                     know
                        0.00
                   america
                        0.00
                           bertha
                           bombing
      NOT insult
                                     insult
               comfortable
0.01
                  designed
                           malignant
                           0.01
                           dries
                 represents
                            undermine
                           never
                       love
                        0.00
                   muslims
                        0.00
                           0.00
NOT identity_attack
                              identity_attack
                           muslims
0.03
                           mob
                           bertha
                           0.01
                    cooper
0.01
                           malignant
               comfortable
                 represents
                        0.01
                  anything
                        0.0
                  designed
NOT sexual_explicit
                               sexual_explicit
                 represents
                           mob
                           bertha
                           0.00
                       iran
                        0.00
```

```
anything
                        0.0
                        mr
                        0.0
                   america
                           malignant
                           move
    NOT obscene
                                  obscene
                represents
               comfortable
                           mob
                           0.00
                           written
                           იიი
                           know
                           0.00
                           bombing
                          dries
                           0.00
                          america
                           0.00
                        mı
                        0.00
                  muslims
NOT severe_toxicity
                              severe_toxicity
                           promote
                  anything
                           lmob
                           0.00
                     neve
                        0.00
                           malignant
                           move
                           represents
                           0.00
                          friends
                           0.00
                          muslims
Text with highlighted words
mr dries bertha cooper represents malignant mob never written anything designed undermine america promote muslims love move iran
wait bombing know comfortable friends
```

Train: Correct Predictions (Strong Case)

LIME analysis :Class 0

```
In[]:
train['predicted'] = train_classes

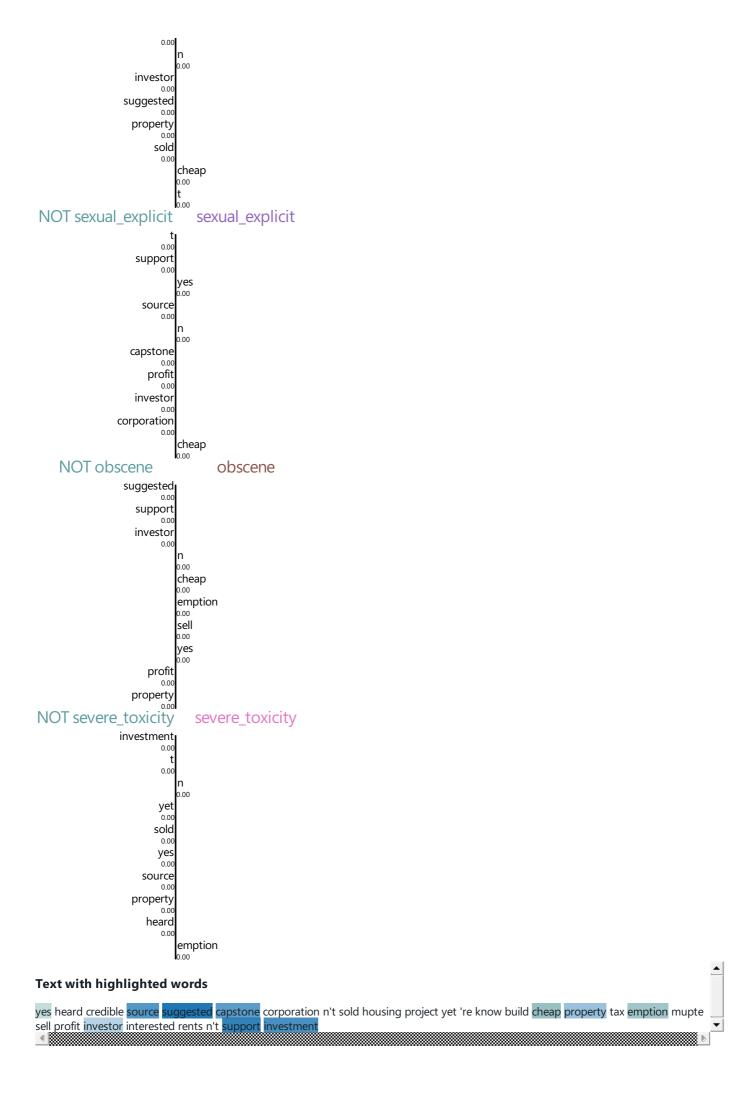
In[]:
new_df = train.loc[train['sub_toxic']==0]
new_df = new_df[new_df['predicted'] == 0]
new_df.reset_index(drop=True, inplace=True)

In[]:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])

Actual class: 0
Predicted class: 0
```

```
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,exp.show in notebook(text=new_df["clean_comment"][idx])
```

Prediction probabilities non-toxic 0.99 insult 0.01 obscene 0.00 sexual_explicit 0.00 Other 0.00 NOT non-toxic non-toxic suggested 0.01 support 0.01 investment 0.01 source 0.00 capstone cheap emption property 0.00 investor **NOT** threat threat investment suggested property 0.0 investor 0.0 support source capstone credible 0.0 interested **NOT** insult insult suggested support investment source 0.0 capstone cheap emption property 0.00 build 0.0 NOT identity_attack identity_attack support 0.00 investment source



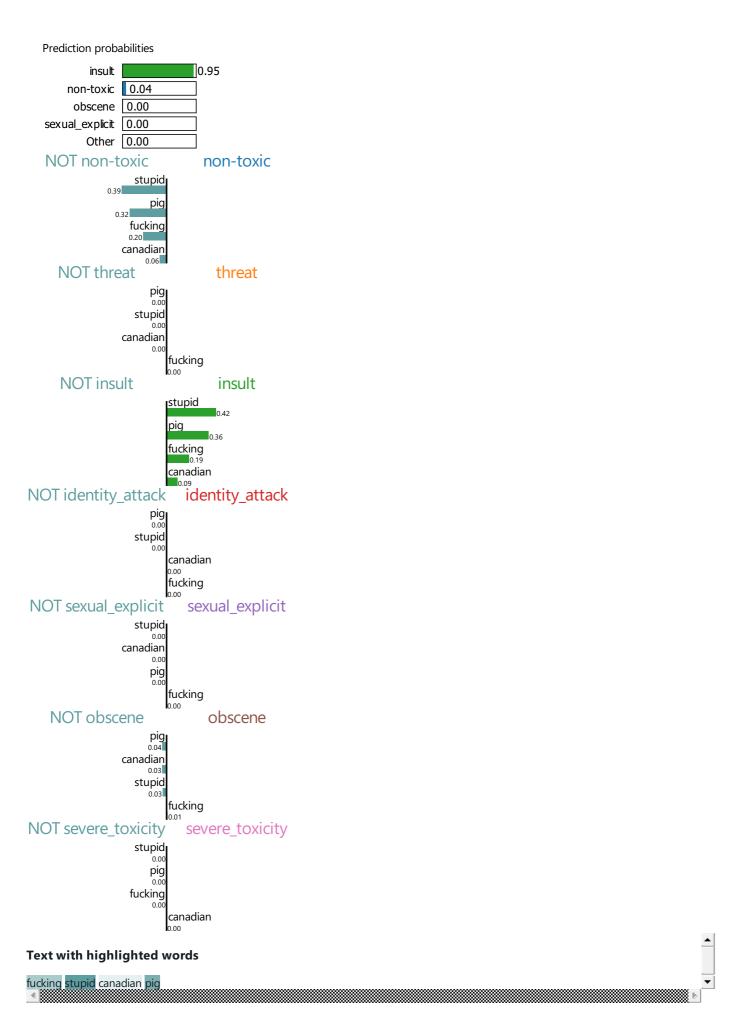
LIME analysis: Class 1

```
In []:
train['predicted'] = train classes
                                                                                                            In []:
new_df = train.loc[train['sub_toxic']==1]
new_df = new_df[new_df['predicted'] == 1]
new_df.reset_index(drop=True, inplace=True)
                                                                                                            In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new df['predicted'][idx])
Actual class: 1
Predicted class: 1
                                                                                                            In [ ]:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
```

```
Prediction probabilities
           threat
                         0.54
                         0.45
       non-toxic
   identity_attack 0.01
            insult 0.01
           Other 0.00
   NOT non-toxic
                                 non-toxic
                    shoot
               0.39
                    guns
<sub>0.14</sub>
                    homes
                          peoples
<sub>0.04</sub>
      NOT threat
                                    threat
                          shoot
                          guns
                          0.14
homes
                           0.07
                   peoples
      NOT insult
                                    insult
                          peoples
                          shoot
                           0.00
                    homes
                       0.0
                     guns
                       0.00
NOT identity_attack
                             identity_attack
                          peoples
                    homes
                           shoot
                          guns
NOT sexual_explicit
                              sexual explicit
                          homes
                           peoples
                           0.00
                          shoot
    NOT obscene
                                  obscene
                          homes
                          0.00
                          shoot
                           0.00
                          peoples
                          0.00
                          guns
<sub>0.00</sub>
NOT severe_toxicity
                              severe_toxicity
                    homes
                          peoples
                           0.00
                          shoot
                          0.00
                          guns
0.00
Text with highlighted words
guns peoples homes shoo
```

LIME analysis :Class 2

exp.show_in_notebook(text=new_df["clean_comment"][idx])



LIME analysis : Class 3

In []:
new_df = train.loc[train['sub_toxic']==3]
new_df = new_df[new_df['predicted'] == 3]
new_df.reset_index(drop=True, inplace=True)

In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])

Actual class: 3
Predicted class: 3
In []:
class_names = ['non-toxic', 'threat', 'insult', 'identity_attack', 'sexual_explicit', 'obscene', 'severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)

exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,

exp.show_in_notebook(text=new_df["clean_comment"][idx])

```
Prediction probabilities
                          0.80
  identity_attack
      non-toxic 0.17
          insult 0.02
         threat 0.00
         Other 0.00
   NOT non-toxic
                             non-toxic
                 muslim
                muslims
               0.30
     NOT threat
                                threat
                       terrorists
                 muslim
                     0.0
                muslims
     NOT insult
                                insult
                       terrorists
                 muslim
                    0.04
                muslims
NOT identity_attack
                          identity_attack
                       muslim
                       terrorists
                       muslims
NOT sexual_explicit
                           sexual_explicit
                       terrorists
                        0.00
                 muslim
                        muslims
    NOT obscene
                              obscene
                       terrorists
                 muslim
                     0.0
                muslims
NOT severe_toxicity
                          severe_toxicity
                       terrorists
                        muslims
                 muslim
Text with highlighted words
muslims terrorists terrorists muslim
LIME analysis: Class 4
                                                                                                                  In []:
train['predicted'] = train classes
                                                                                                                  In []:
new_df = train.loc[train['sub_toxic']==4]
new_df = new_df[new_df['predicted'] == 4]
new_df.reset_index(drop=True, inplace=True)
                                                                                                                  In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
```

```
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,exp.show_in_notebook(text=new_df["clean_comment"][idx])
```

Prediction probabilities sexual_explicit 0.38 non-toxic insult 0.02 identity_attack 0.01 Other 0.00 NOT non-toxic non-toxic penis vagina understand problem 0.09 use 0.06 wanting restroom NOT threat threat wanting penis 0.00 use 0.0 problem vagina understand restroom **NOT** insult insult vagina restroom problem wanting understand NOT identity_attack identity_attack vagina restroom use 0.00 penis 0.00 problem understand wanting _{0.00} NOT sexual_explicit sexual_explicit penis vagina 0.22

```
understand
                 problem
                   0.10
                     use
                    0.07
                        wanting
                         0.06
                        restroom
    NOT obscene
                               obscene
                     use
                 wanting
                     0.0
                        restroom
                  vagina
                     0.0
                   penis
                     0.0
                        understand
NOT severe_toxicity
                           severe_toxicity
                 wanting
                     0.0
                     use
                     0.0
                  vagina
                     0.0
                        understand
                 problem
                     0.0
                restroom
Text with highlighted words
problem penis wanting use vagina restroom vagina wanting use penis restroom understand
LIME analysis : Class 5
                                                                                                                     In []:
train['predicted'] = train_classes
                                                                                                                      In []:
new_df = train.loc[train['sub_toxic']==5]
new_df = new_df[new_df['predicted'] == 5]
new_df.reset_index(drop=True, inplace=True)
                                                                                                                     In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
```

class names = ['non-toxic','threat','insult','identity attack','sexual explicit','obscene','severe toxicity']

exp = explainer.explain instance(new df["clean comment"][idx], bert.predict, num features = 10,labels=[0, 1,2,

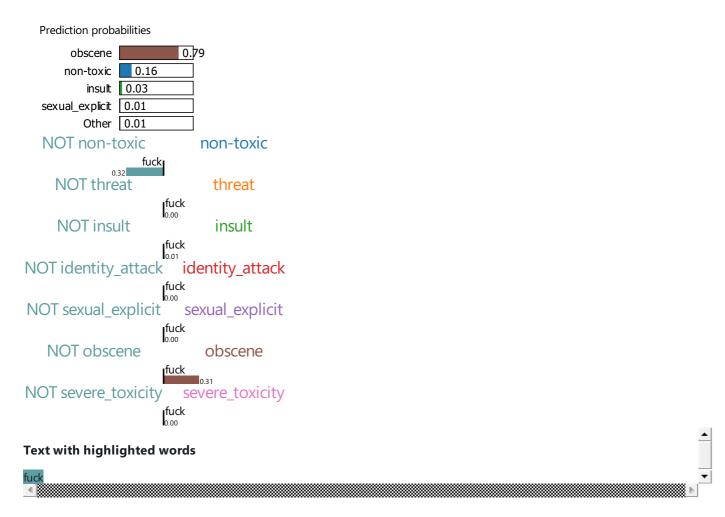
In []:

print('Predicted class: ', new_df['predicted'][idx])

explainer = LimeTextExplainer(class names = class names)

exp.show_in_notebook(text=new_df["clean_comment"][idx])

Actual class: 5
Predicted class: 5

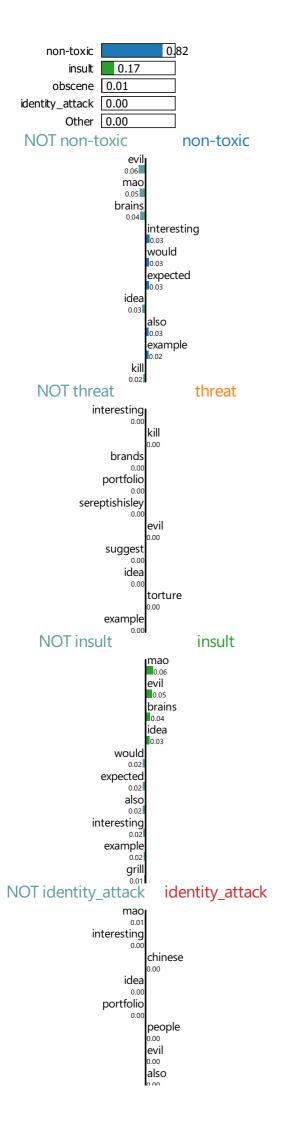


Test: Incorrect predictions (Worst Case)

LIME analysis: Class 0

```
In []:
test['predicted'] = test_classes
                                                                                                            In []:
new df = test.loc[test['sub toxic']==0]
new_df = new_df[new_df['predicted'] != 0]
new_df.reset_index(drop=True, inplace=True)
                                                                                                            In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 0
Predicted class: 2
                                                                                                            In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
```

```
Prediction probabilities
                       0.65
          insult
                   0.33
      non-toxic
       obscene 0.02
         threat 0.00
         Other 0.00
   NOT non-toxic
                            non-toxic
           gobbledygook
                 bunch
     NOT threat
                              threat
                      bunch
          gobbledygook
     NOT insult
                               insult
                      gobbledygook
0.19
                      bunch
NOT identity_attack
                         identity_attack
                 bunch
          gobbledygook
NOT sexual_explicit
                          sexual_explicit
                      gobbledygook
   NOT obscene
                             obscene
                      bunch
                      gobbledygook
NOT severe toxicity
                 bunch
                       gobbledygook
Text with highlighted words
bunch gobbledygook
LIME analysis : Class 1
                                                                                                             In []:
test['predicted'] = test classes
                                                                                                             In []:
new_df = test.loc[test['sub_toxic']==1]
new_df = new_df[new_df['predicted'] != 1]
new_df.reset_index(drop=True, inplace=True)
                                                                                                             In [ ]:
idx= np.random.random integers(0,len(new df))
print('Actual class: ', new df['sub toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 1
Predicted class: 0
                                                                                                             In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
```



```
bar
NOT sexual_explicit
                             sexual_explicit
                          brains
                          suggest
                          torture
               interesting
                    good
                  chinese
                  example
    NOT obscene
                                 obscene
                         brains
                  suggest
                          grill
                    would
               interesting
                  chinese
                      0.00
                      also
                  example
                       0.00
                     idea
NOT severe toxicity
                             severe_toxicity
                     idea
                      0.0
                     mao
                       0.00
                    good
                          also
                wholesale
                          suggest
                          example
                          0.00
                          would
                          mean
                          0.00
                         brains
```

Text with highlighted words

chinese people kill cigarette smokers n't know torture like guantanamo james sayeth smith kline french n't portfolio pharmacy fred meyer patrons joes bar grill wholesale slaughter also mean perpetrating evil eyes really expected switch brands would suggest sereptishisley drank alcohol smoked ciggies one blow brains set good example interesting idea mao

In []:

In []:

LIME analysis : Class 2

```
test['predicted'] = test_classes
new_df = test.loc[test['sub_toxic']==2]
```

```
new df = new df[new df['predicted'] != 2]
new df.reset index(drop=True, inplace=True)
                                                                                                                    In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new df['sub toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 2
Predicted class: 0
                                                                                                                    In []:
class names = ['non-toxic','threat','insult','identity attack','sexual explicit','obscene','severe toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10, labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
  Prediction probabilities
                          0.75
       non-toxic
          insult
                  0.21
  identity_attack 0.03
       obscene 0.00
         Other 0.00
   NOT non-toxic
                              non-toxic
                  0.16
              misogynist
                        republicans
                  group
                    0.06
                 bigoted
                    0.05
                        party
                        bulbs
                        republican
                        0.03
                    lots
                    0.03
                        watch
     NOT threat
                                threat
                       suicide
                        0.00
                   party
              republicans
                  today
                 bigoted
                     0.0
                   knew
                     0.0
              misogynist
                     0.0
                  group
                     0.0
                    dim
     NOT insult
                                insult
                       racist
                        misogynist
              republicans
                        bigoted
                        0.06
```

group

bulbs 0.04 party

```
lots
                        0.02
                        left
NOT identity_attack
                              identity_attack
                     today
0.02
                            0.02
                            misogynist
                           suicide
                republicans
                       dim
                        0.01
                 republican
                   bigoted
                      party
NOT sexual_explicit
                               sexual_explicit
                 republican
                            misogynist
                republicans
                            group
                            lots
                     watch
                        0.00
                            knew
                        left
0.00
                     today
    NOT obscene
                                   obscene
                           knew
                            today
                           left
                            0.00
                 republican
                            bigoted
                      party
                    suicide
                           republicans
                           <sub>0.00</sub>
bulbs
NOT severe_toxicity
                              severe_toxicity
                     knew
0.00
lots
                   bigoted
                           racist
                 republican
                            suicide
                republicans
                           group
```

Text with highlighted words

lots republicans suicide watch today republican party knew group racist misogynist bigoted dim bulbs left

LIME analysis :Class 3

```
In []:
test['predicted'] = test_classes
                                                                                                                    In []:
new df = test.loc[test['sub toxic']==3]
new_df = new_df[new_df['predicted'] != 3]
new_df.reset_index(drop=True, inplace=True)
                                                                                                                    In [ ]:
idx= np.random.random integers(0,len(new df))
print('Actual class: ', new df['sub toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 3
Predicted class: 0
                                                                                                                    In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10, labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
  Prediction probabilities
                            0.90
      non-toxic
  identity_attack 0.07
          insult 0.03
       obscene 0.00
          Other 0.00
   NOT non-toxic
                              non-toxic
                       must
                        0.05
                        comes
                        0.04
                 muslim
                    0.03
                   expe
                    0.03
                  crime
                    0.02
                        religion
                 terrible
                     0.02
                   know
                    rubs
                     0.02
                   islam
                     0.01
     NOT threat
                                threat
                 religion
                existence
                     0.0
                  know
              considered
                believers
                     0.0
                        death
                        0.00
                        lexpel
```

```
crime
                            0.00
                            terrible
      NOT insult
                                      insult
                      must
                            terrible
                            expel
                    muslim
                    religion
0.01
islam
                        0.01
                    atrocity
                            crime
                            mind
                           rubs
NOT identity_attack
                               identity_attack
                           muslim
0.05
                      must
                       0.04
                     comes
                        0.03
                            islam
                            0.02
                            know
                            0.02
                            crime
                  cherished
                  applauds
                           rubs
<sub>0.01</sub>
NOT sexual_explicit
                               sexual_explicit
                      must
0.00
                 considered
                        still
                        0.00
                     comes
                  humanity
                            know
                            0.00
                            hands
    NOT obscene
                                    obscene
                      must
0.00
                            religion
                  intolerant
                            crime
                            rubs
                    muslim
                        0.00
                     obtain
                            know
                            death
```

```
NOT severe_toxicity

religion
considered
con
```

Text with highlighted words

muslim causes terrible atrocity world still rubs hands islam religion considered one obtain religious freedom cherished west intolerant way life applauds death non believers world wide organized crime syndicate comes mind guise religion humanity must expel existence know

LIME analysis: Class 4

predator 0.01 mean 0.01 giving

threat

NOT threat

```
In []:
test['predicted'] = test classes
                                                                                                               In []:
new df = test.loc[test['sub toxic']==4]
new df = new df[new df['predicted'] != 4]
new df.reset index(drop=True, inplace=True)
                                                                                                               In []:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 4
Predicted class:
                                                                                                               In []:
class names = ['non-toxic','threat','insult','identity attack','sexual explicit','obscene','severe toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
  Prediction probabilities
      non-toxic
                           0.93
          insult 0.03
  sexual_explicit 0.03
  identity_attack 0.01
         Other 0.00
   NOT non-toxic
                            non-toxic
                 sexual
                       lost
              billionaire
                       women
```

```
women
                        0.0
                            predator
                  billionaire
                           giving
<sub>0.00</sub>
                           mean
                            0.00
                           sexual
      NOT insult
                                     insult
                    women
                           billionaire
                            0.03
                           mean
                           0.02
                           predator
0.01
                           sexual
                           0.01
                        0.01
                     giving
NOT identity_attack
                              identity_attack
                           sexual
                  billionaire
                        0.01
                     mear
                  predato
                     giving
NOT sexual_explicit
                               sexual_explicit
                           giving
                     mean
0.01
                           billionaire
                           0.01
                           predator
    NOT obscene
                                   obscene
                           billionaire
                           predator
                            0.00
                           sexual
                            0.00
                   women
NOT severe_toxicity
                               severe_toxicity
                     giving
                     mean
                        0.0
                  billionaire
                        0.0
                    women
                        0.00
                            sexual
                       lost
                        0.00
                           Inredator
```

0.00

Text with highlighted words

sexual predator mean women giving billionaire women lost

LIME analysis : Class 5

```
In []:
test['predicted'] = test classes
                                                                                                               In []:
new_df = test.loc[test['sub_toxic']==5]
new_df = new_df[new_df['predicted'] != 5]
new_df.reset_index(drop=True, inplace=True)
                                                                                                               In []:
idx= np.random.random integers(0,len(new df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 5
Predicted class: 0
                                                                                                               In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class names = class names)
exp = explainer.explain instance(new df["clean comment"][idx], bert.predict, num features = 10,labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
  Prediction probabilities
      non-toxic
                           0.96
          insult 0.03
       obscene 0.01
         threat 0.00
         Other 0.00
   NOT non-toxic
                            non-toxic
                 subaru
                    0.03
                       lname
                       0.02
                   mis
                       wtf
                       parent
                   spell
                               threat
```

NOT threat parent spell subaru name 0.0 wtf

NOT insult insult

> subaru name 0.02 wt mis 0.01 spell 0.01 parent

```
NOT identity_attack
                            identity_attack
                      wtf
                   parent
                      0.0
                     spel
                      0.0
                         name
                  subaru
NOT sexual_explicit
                            sexual_explicit
                   parent
                         subaru
                         0.00
                   name
                      0.0
                         mis
                     spell
                      0.0
                      wtf
    NOT obscene
                                obscene
                         subaru
                         wtf
                         0.00
                   name
                      0.0
                     spell
                      0.00
                     mis
                   parent
NOT severe_toxicity
                            severe_toxicity
                   name
                   parent
                         mis
                         0.00
                    spell
Text with highlighted words
wtf name parent mis-spell subaru
LIME analysis : Class 6
```

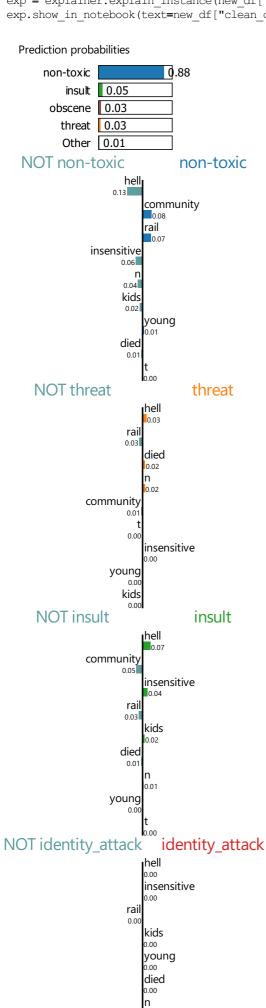
```
ln[]:
test['predicted'] = test_classes

ln[]:
new_df = test.loc[test['sub_toxic']==6]
new_df = new_df[new_df['predicted'] != 6]
new_df.reset_index(drop=True, inplace=True)

ln[]:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])

Actual class: 6
Predicted class: 0

ln[]:
class_names = ['non-toxic', 'threat', 'insult', 'identity_attack', 'sexual_explicit', 'obscene', 'severe_toxicity']
```



0.00

```
COMMUNIC
NOT sexual_explicit
                              sexual_explicit
                           hell
                           insensitive
                community
                           kids
                           0.00
                           0.00
                           0.00
                           young
                           0.00
    NOT obscene
                                  obscene
                           hell
                           0.01
                    young
                       0.0
                       rail
                       0.01
                community
                           insensitive
                           kids
                           0.01
                           o.oo
died
NOT severe_toxicity
                              severe_toxicity
                           died
                           community
                        0.0
                           hell
                       kids
                           insensitive
                    young
                        0.0
                        rail
Text with highlighted words
n't insensitive young kids community died hell rail
```

Test: Correct Predictions (Strong Case)

LIME analysis : Class 0

```
In []:
test['predicted'] = test_classes

In []:
new_df = test.loc[test['sub_toxic']==0]
new_df = new_df[new_df['predicted'] == 0]
new_df.reset_index(drop=True, inplace=True)
```

In []:

```
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 0
Predicted class: 0
                                                                                                                     In []:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain instance(new df["clean comment"][idx], bert.predict, num features = 10,labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
  Prediction probabilities
      non-toxic
                              0.99
          insult 0.00
  identity_attack 0.00
         threat 0.00
          Other 0.00
   NOT non-toxic
                              non-toxic
                        article
                        0.01
                        report
                        0.00
                   hang
                        period
                        physicians
                        explains
                        treatment
                        0.00
                    beat
                        center
                 digging
     NOT threat
                                threat
                  article
                  report
                     0.00
                        hang
               physicians
                  period
                        closed
                 remains
                     0.0
                  center
                     0.0
                        four
                        0.00
                        leave
     NOT insult
                                 insult
               treatment
                     0.0
                  center
                     0.0
                 explains
                     0.0
               physicians
```

period

report

beat 0.00 digging 0.00

hang

```
happened
0.00
NOT identity_attack
                               identity_attack
                            peoples
0.00
                     cente
                         0.00
                     report
                         0.00
                      needs
                             lives
                             0.00
                     period
                         0.00
                            closed
                             0.00
                            happen
                             0.00
                            beat
                    balance
NOT sexual_explicit
                                sexual_explicit
                   remains
                      article
                         0.00
                     report
                         0.0
                            digging
                    explains
                            hang
                      follow
                         0.0
                             treatment
                      short
                         0.00
                        civil
    NOT obscene
                                    obscene
                    explains
                    remains
                         0.00
                  physicians
                             hang
                     period
                            digging
<sub>0.00</sub>
                             center
                  treatment
                         0.00
                     report
                         0.00
                            beat
NOT severe_toxicity
                               severe_toxicity
                            center
                      follow
                         0.0
                       time
                         0.00
                             lives
                    remains
                            peoples
                             0.00
                            closed
                    balance
                         0.00
                      needs
                         0.00
                  happened
```

LIME analysis : Class 1

0.14 crime 0.11 committing 0.03 pull ก กว

```
In [198]:
test['predicted'] = test_classes
                                                                                                                  In [199]:
new df = test.loc[test['sub toxic']==1]
new df = new df[new df['predicted'] == 1]
new df.reset index(drop=True, inplace=True)
                                                                                                                  In [200]:
new_df
                                                                                                                  Out[200]:
       id comment_text split
                                created_date publication_id parent_id article_id
                                                                            rating funny wow sad likes disagree toxicity sev
             If you pull a
               gun while
                                  2016-08-21
  7085319
                                                                   144243 approved
                                                                                                             0 0.581081
                                                     21
                                                            NaN
                                                                                      0
                                                                                           0
                                                                                               0
                        test
            committing a
                            21:45:56.822437+00
             crime you...
                                                                                                                  In [201]:
idx = 0
                                            #because only 1 index
print('Actual class: ', new df['sub toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 1
Predicted class: 1
                                                                                                                   In [202]:
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10, labels=[0, 1,2,
exp.show in notebook(text=new df["clean comment"][idx])
  Prediction probabilities
                           0.79
          threat
       non-toxic 0.20
          insult 0.00
       obscene 0.00
          Other 0.00
   NOT non-toxic
                              non-toxic
                    shot
                 deserve
                 0.18
                    gun
                  0.14
                   crime
                   0.13
              committing
                    0.04
                    pull
     NOT threat
                                threat
                        shot
                                  0.53
                        deserve
                          0.16
                        gun
```

```
NOT insult
                                    insult
                      shot
                          deserve
                          0.02
                          committing
                          0.01
                          crime
                          0.00
                          pull
                          0.00
NOT identity_attack
                             identity_attack
                          crime
                          0.00
                          committing
                          deserve
                          0.00
                          gun
                          pull
NOT sexual_explicit
                              sexual_explicit
                      shot
                       0.00
                          committing
                          0.00
                          pull
                          0.00
                          deserve
                          crime
                          0.00
                      gun
0.00
    NOT obscene
                                  obscene
                      shot
0.01
                          deserve
                          0.00
                          pull
                          crime
                          0.00
                          gun
                          committing
NOT severe_toxicity
                             severe_toxicity
                     shot
0.00
                          crime
                          0.00
                          committing
                          pull
                          deserve
                          0.00
                          gun
0.00
Text with highlighted words
pull gun committing crime deserve shot
```

LIME analysis: Class 2

```
test['predicted'] = test classes
                                                                                                          In [195]:
new_df = test.loc[test['sub_toxic']==2]
new_df = new_df[new_df['predicted'] == 2]
new_df.reset_index(drop=True, inplace=True)
```

In [194]:

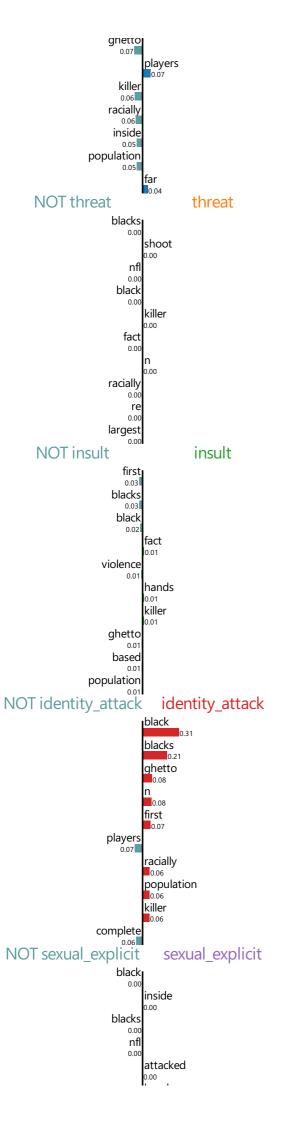
In [197]:

```
idx= np.random.random integers(0,len(new df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new_df['predicted'][idx])
Actual class: 2
Predicted class: 2
class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
explainer = LimeTextExplainer(class names = class names)
exp = explainer.explain instance(new df["clean comment"][idx], bert.predict, num features = 10,labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
  Prediction probabilities
          insult
                       0.56
      non-toxic
                      0.44
          threat 0.00
       obscene 0.00
          Other 0.00
   NOT non-toxic
                              non-toxic
                stupidity
                        persists
                        <u>ev</u>erything
                  world
                   0.08
                    rest
                    0.07
                  trump
     NOT threat
                                threat
                 persists
                        everything
                        0.00
                        world
                stupidity
                  trump
                     0.0
                    rest
     NOT insult
                                insult
                        stupidity
                 persists
                   0.10
              everything
                   0.10
                         0.08
                        rest
                         0.07
                       trump
identity_attack
NOT identity_attack
                        world
                        າ ດດ
              everything
                     0.0
                stupidity
                  trump
                     0.0
                 persists
                     0.0
                    res
```

NOT sexual explicit

world 0.0 everything persists sexual explicit

```
trump
                     stupidity
   NOT obscene
                           obscene
               persists
                     everything
                world
                   0.00
                trump
                     stupidity
NOT severe toxicity
                        severe toxicity
                world
                   0.0
             everything
               persists
                     trump
                     rest
Text with highlighted words
trump rest world everything persists stupidity
LIME analysis : Class 3
                                                                                                    In [176]:
test['predicted'] = test classes
                                                                                                    In [177]:
new df = test.loc[test['sub toxic']==3]
new df = new df[new df['predicted'] == 3]
new df.reset index(drop=True, inplace=True)
                                                                                                    In [178]:
idx= np.random.random_integers(0,len(new_df))
print('Actual class: ', new_df['sub_toxic'][idx])
print('Predicted class: ', new df['predicted'][idx])
Actual class: 3
Predicted class: 3
                                                                                                    In [179]:
class names = ['non-toxic','threat','insult','identity attack','sexual explicit','obscene','severe toxicity']
explainer = LimeTextExplainer(class_names = class_names)
exp = explainer.explain_instance(new_df["clean_comment"][idx], bert.predict, num_features = 10,labels=[0, 1,2,
exp.show_in_notebook(text=new_df["clean_comment"][idx])
  Prediction probabilities
  identity_attack
                     0.62
                  0.35
      non-toxic
         insult 0.03
       obscene 0.00
         Other 0.00
   NOT non-toxic
                          non-toxic
                blacks
                 0.09
```



```
hands
                             0.00
                             0.00
                             racially
                     change
                      based
    NOT obscene
                                     obscene
                      blacks
                       black
                          0.0
                             lfact
                             0.00
                             0.00
                 population
                             touch
                             0.00
                             killer
                             0.00
                             complete
                             0.00
                             inside
NOT severe_toxicity
                                severe_toxicity
                      blacks
                             touch
                             0.00
                        fact
                             killer
                             0.00
                             0.00
                             violence
                             shoot
                             0.00
                             protesting
                             0.00
                             enforcement
                             0.00
                          0.00
Text with highlighted words
first 're protesting based lies blacks racially attacked law enforcement hands n't shoot complete lie nfl wants help black population need
change ghetto culture inside black violence far largest killer blacks nfl players n't touch fact
```

LIME analysis :Class 5

```
In [186]:

test['predicted'] = test_classes

In [187]:

new_df = test.loc[test['sub_toxic']==5]
 new_df = new_df[new_df['predicted'] == 5]
 new_df.reset_index(drop=True, inplace=True)

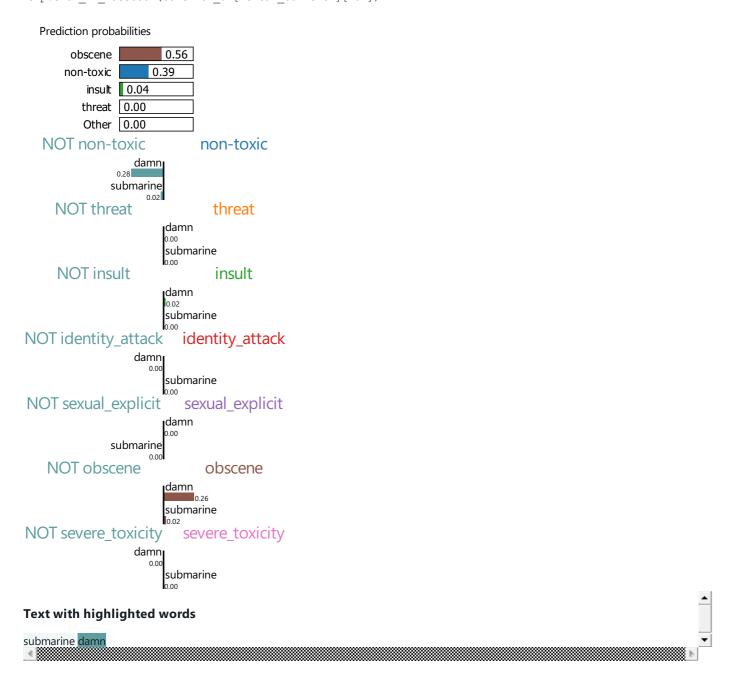
In [188]:

idx= np.random.random_integers(0,len(new_df))
 print('Actual class: ', new_df['sub_toxic'][idx])
 print('Predicted class: ', new_df['predicted'][idx])

Actual class: 5
 Predicted class: 5

In [189]:

class_names = ['non-toxic','threat','insult','identity_attack','sexual_explicit','obscene','severe_toxicity']
 explainer = LimeTextExplainer(class_names = class_names)
```



Points to be noted:-

- Class 0 non-toxic:-
 - Train It is doing decent job in learning. Because of some words like murderous and traitor it incorrectly classified.
 - Test It is doing good in classification. It did wrong classification because of some made up word.
- Class 1 threat:-
 - Test and Train It is not doing that bad as the words because of which it is being classified as a threat are normal words and in the correct classification, we can see it is learning well to classify those words as threads.
- Class 2 insult:-
 - Train and Train It is identifying well the words that belong to insult category but because of presence of more non-toxic words it is classifying them as non-toxic.
- Class 3 identity attack:-
 - Train and Test It is not giving weightage to identity words like islam and white. It is doing well to classify identity attack on blacks.
- Class 4 threat:-
 - Train and Test It is not giving weightage to explicit words like naked, predator. It is doing well in knowing that they are toxic but due to presence of lots of non-toxic words it is classifying them as non-toxic.
- Class 5 obscene:-
 - Train and Test It is doing well in identifying the obscene words like fuck but it is not classifying other obscene words like rascist, crap well.
- Class 6 severe_toxicity:-
 - Train and Test There are not enough examples for it to learn.

(Not included LIME analysis for any of the class in the sub-category of test or train where no data point was available)

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js