t-analysis-and-offensive-languages

February 19, 2024

1 Offensive Language Identiication Task

- StackOverflow
- Word2Vec Gensim
- Embedding Layer

[1]:
```
import pandas as pd
import re


import nltk
from nltk import word_tokenize, pos_tag
from nltk.collocations import *
from itertools import *
from nltk.util import ngrams
from nltk.corpus import stopwords
```

[2]:
```
data = pd.read_csv("labeled_data.csv")
data
```

[2]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	\
0	0	3	0	0	3	2	
1	1	3	0	3	0	1	
2	2	3	0	3	0	1	
3	3	3	0	2	1	1	
4	4	6	0	6	0	1	
…	…	…	…	…	…	…	
24778	25291	3	0	2	1	1	
24779	25292	3	0	1	2	2	
24780	25294	3	0	3	0	1	
24781	25295	6	0	6	0	1	
24782	25296	3	0	0	3	2	

	tweet
0	!!! RT @mayasolovely: As a woman you shouldn't…
1	!!!!! RT @mleew17: boy dats cold…tyga dwn ba…
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby…
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo…

```
4      !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you…
…                                                    …
24778  you's a muthaf***in lie &#8220;@LifeAsKing: @2…
24779  you've gone and broke the wrong heart baby, an…
24780  young buck wanna eat!!.. dat nigguh like I ain…
24781              youu got wild bitches tellin you lies
24782  ~~Ruffled | Ntac Eileen Dahlia - Beautiful col…

[24783 rows x 7 columns]
```

[3]:
```
data.drop('Unnamed: 0', axis = 1, inplace = True)
```

[4]:
```
data.columns
```

[4]:
```
Index(['count', 'hate_speech', 'offensive_language', 'neither', 'class',
       'tweet'],
      dtype='object')
```

[5]:
```
rege = r"[^a-zA-Z0-9 ]"
data['tweet'] = data['tweet'].str.replace('@\S+', '', regex=True)
data['tweet'] = data['tweet'].str.replace(rege, '', regex=True)
data['tweet'] = data['tweet'].str.replace("RT", '')
data
```

[5]:

	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	As a woman you shouldnt complain about clea…
1	3	0	3	0	1	boy dats coldtyga dwn bad for cuffin dat ho…
2	3	0	3	0	1	Dawg You ever fuck a bitch and she start …
3	3	0	2	1	1	she look like a tranny
4	6	0	6	0	1	The shit you hear about me might be true or…
…	…	…	…	…	…	…
24778	3	0	2	1	1	yous a muthafin lie 8220 right His TL is tra…
24779	3	0	1	2	2	youve gone and broke the wrong heart baby and …
24780	3	0	3	0	1	young buck wanna eat dat nigguh like I aint fu…
24781	6	0	6	0	1	
24782	3	0	0	3	2	

```
24781              youu got wild bitches tellin you lies
24782  Ruffled  Ntac Eileen Dahlia  Beautiful color c…

[24783 rows x 6 columns]
```

[6]:
```
# Stopwords removal and lowercase
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

data['tweet'] = data['tweet'].str.lower()
data['tweet'] = data['tweet'].apply(lambda x: ' '.join([word for word in x.
 ↪split() if word not in (stop_words)]))

data
```

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

[6]:
```
       count  hate_speech  offensive_language  neither  class  \
0          3            0                   0        3      2
1          3            0                   3        0      1
2          3            0                   3        0      1
3          3            0                   2        1      1
4          6            0                   6        0      1
…      …           …                   …       …    …
24778      3            0                   2        1      1
24779      3            0                   1        2      2
24780      3            0                   3        0      1
24781      6            0                   6        0      1
24782      3            0                   0        3      2

                                                   tweet
0      woman shouldnt complain cleaning house amp man…
1      boy dats coldtyga dwn bad cuffin dat hoe 1st p…
2           dawg ever fuck bitch start cry confused shit
3                                      look like tranny
4      shit hear might true might faker bitch told ya…
…                                                    …
24778  yous muthafin lie 8220 right tl trash 8230 min…
24779  youve gone broke wrong heart baby drove rednec…
24780  young buck wanna eat dat nigguh like aint fuck…
24781                  youu got wild bitches tellin lies
24782  ruffled ntac eileen dahlia beautiful color com…

[24783 rows x 6 columns]
```

[7]:
```
# Remove numbers
data['tweet'] = data['tweet'].str.replace('\d', '', regex=True)

data
```

[7]:

	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	woman shouldnt complain cleaning house amp man…
1	3	0	3	0	1	boy dats coldtyga dwn bad cuffin dat hoe st place
2	3	0	3	0	1	dawg ever fuck bitch start cry confused shit
3	3	0	2	1	1	look like tranny
4	6	0	6	0	1	shit hear might true might faker bitch told ya
…	…	…	…	…	…	…
24778	3	0	2	1	1	yous muthafin lie right tl trash mine bible …
24779	3	0	1	2	2	youve gone broke wrong heart baby drove rednec…
24780	3	0	3	0	1	young buck wanna eat dat nigguh like aint fuck…
24781	6	0	6	0	1	youu got wild bitches tellin lies
24782	3	0	0	3	2	ruffled ntac eileen dahlia beautiful color com…

[24783 rows x 6 columns]

2 Word Emeddings

1. Word2Vec
2. FastText
3. CNN
4. RNN

[8]:
```
# from gensim.test.utils import common_texts
from gensim.models import Word2Vec


# model = Word2Vec(sentences=common_texts, vector_size=100, window=5,␣
 ↪min_count=1, workers=4)
# model.save("word2vec.model")
```

[9]:
```
# model = Word2Vec.load("word2vec.model")
# model.train(data["tweet"], total_examples=model.corpus_count, epochs=model.
 ↪epochs)
```

[10]:
```
word2vec_model = Word2Vec(data['tweet'], min_count=1, vector_size=100, window=5)
```

```
WARNING:gensim.models.word2vec:Each 'sentences' item should be a list of words
(usually unicode strings). First item here is instead plain <class 'str'>.
```

[11]:
```
word2vec_model.wv[1]
```

[11]:
```
array([-0.22526443,  0.25804466,  0.2666585 ,  0.41574985, -0.03260519,
       -0.28431383, -0.17225118, -0.21790202, -0.07212105, -0.09368566,
       -0.01206284, -0.12150647, -0.02750993,  0.21436875,  0.31730402,
        0.3334212 ,  0.01582862,  0.14916794, -0.18448913, -0.0280059 ,
        0.20831497,  0.12793839,  0.16162473, -0.2605099 ,  0.2916233 ,
        0.2445141 , -0.43050373, -0.20226139,  0.00481574,  0.12548327,
        0.15627855, -0.0904948 ,  0.1568621 ,  0.01202086,  0.3338502 ,
        0.4517225 ,  0.08700998,  0.26569566,  0.17754984,  0.10538246,
        0.21800874, -0.3668148 , -0.27247864,  0.07558532, -0.26101997,
        0.01794637, -0.09543937, -0.13037425,  0.10369914, -0.01893698,
        0.09546362, -0.32084873, -0.16343908,  0.30054814,  0.19292213,
        0.47139403,  0.25672436,  0.1558889 , -0.0423195 ,  0.40952352,
       -0.24772684,  0.3369923 , -0.25124922, -0.10704596,  0.3229313 ,
       -0.00796977,  0.08009604, -0.1578472 ,  0.5580845 , -0.05139901,
       -0.18153004, -0.4315614 , -0.13888766,  0.2725584 ,  0.19501819,
       -0.03961842, -0.41994914, -0.13881381,  0.20255776, -0.10066935,
       -0.18947904,  0.2634208 ,  0.32156816, -0.04277584, -0.04183633,
        0.3141695 ,  0.10460731, -0.24785516,  0.21212351, -0.02401476,
       -0.02196068,  0.12150849,  0.01087661, -0.15318531, -0.25274202,
        0.15040348,  0.03897153, -0.19871005, -0.5026415 ,  0.16289419],
      dtype=float32)
```

[12]:
```
from keras.preprocessing.text import Tokenizer
from keras.utils import pad_sequences
from keras.models import Sequential
from keras.layers import Dense, Flatten, Conv1D, MaxPooling1D, Dropout,␣
 ↪SimpleRNN
from keras.layers import Embedding
```

[13]:
```
df = data.copy()

tokenizer = Tokenizer()
tokenizer.fit_on_texts(df['tweet'])
sequences = tokenizer.texts_to_sequences(df['tweet'])
data = pad_sequences(sequences, maxlen=300)
```

```
model_cnn = Sequential()
model_cnn.add(Conv1D(64, 5, activation='relu', input_shape=(300,100)))
model_cnn.add(MaxPooling1D(pool_size=4))
model_cnn.add(Flatten())
model_cnn.add(Dense(1, activation='sigmoid'))
model_cnn.compile(loss='binary_crossentropy', optimizer='adam',␣
 ↪metrics=['accuracy'])
```

[14]:
```
model_rnn = Sequential()
model_rnn.add(Embedding(len(tokenizer.word_index) + 1, 100, input_length=300))
model_rnn.add(SimpleRNN(100, activation='relu'))
model_rnn.add(Dense(1, activation='sigmoid'))
model_rnn.compile(loss='binary_crossentropy', optimizer='adam',␣
 ↪metrics=['accuracy'])
```