8fcd5a04-40fb-4260-89f9-ef0366c38900 (1)

June 10, 2021

1 « »

- « » . , - . . , .

. .

F1 0.75.

1. .
2. .
3. .

BERT , .

`toxic_comments.csv`. *text* , *toxic* — .

1.1

[1]:
```
!pip install catboost
```

```
Requirement already satisfied: catboost in c:\users\saidd\anaconda3\lib\site-
packages (0.25.1)
Requirement already satisfied: pandas>=0.24.0 in
c:\users\saidd\anaconda3\lib\site-packages (from catboost) (1.0.5)
Requirement already satisfied: six in c:\users\saidd\anaconda3\lib\site-packages
(from catboost) (1.15.0)
```

```
Requirement already satisfied: graphviz in c:\users\saidd\anaconda3\lib\site-
packages (from catboost) (0.16)
Requirement already satisfied: numpy>=1.16.0 in
c:\users\saidd\anaconda3\lib\site-packages (from catboost) (1.18.5)
Requirement already satisfied: scipy in c:\users\saidd\anaconda3\lib\site-
packages (from catboost) (1.5.0)
Requirement already satisfied: plotly in c:\users\saidd\anaconda3\lib\site-
packages (from catboost) (4.9.0)
Requirement already satisfied: matplotlib in c:\users\saidd\anaconda3\lib\site-
packages (from catboost) (3.2.2)
Requirement already satisfied: pytz>=2017.2 in
c:\users\saidd\anaconda3\lib\site-packages (from pandas>=0.24.0->catboost)
(2020.1)
Requirement already satisfied: python-dateutil>=2.6.1 in
c:\users\saidd\anaconda3\lib\site-packages (from pandas>=0.24.0->catboost)
(2.8.1)
Requirement already satisfied: retrying>=1.3.3 in
c:\users\saidd\anaconda3\lib\site-packages (from plotly->catboost) (1.3.3)
Requirement already satisfied: cycler>=0.10 in
c:\users\saidd\anaconda3\lib\site-packages (from matplotlib->catboost) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
c:\users\saidd\anaconda3\lib\site-packages (from matplotlib->catboost) (1.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
c:\users\saidd\anaconda3\lib\site-packages (from matplotlib->catboost) (2.4.7)
```

[2]:

```
#%pip install ipykernel
```

[3]:

```
#                 ,                       )
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np
from string import punctuation
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import SGDClassifier
from sklearn.metrics import f1_score
```

[4]:

```
data = pd.read_csv('toxic_comments.csv', error_bad_lines=False, engine="python")
data.head()
```

[4]:

```
                                                text  toxic
0  Explanation\nWhy the edits made under my usern…      0
1  D'aww! He matches this background colour I'm s…      0
2  Hey man, I'm really not trying to edit war. It…      0
3  "\nMore\nI can't make any real suggestions on …      0
4  You, sir, are my hero. Any chance you remember…      0
```

[5]:
```
try:
    data = pd.read_csv('toxic_comments.csv', error_bad_lines=False,␣
 ↪engine="python")

except:
    data = pd.read_csv('/datasets/toxic_comments.csv', error_bad_lines=False,␣
 ↪engine="python")
data.head()
```

[5]:
```
                                                text  toxic
0  Explanation\nWhy the edits made under my usern…      0
1  D'aww! He matches this background colour I'm s…      0
2  Hey man, I'm really not trying to edit war. It…      0
3  "\nMore\nI can't make any real suggestions on …      0
4  You, sir, are my hero. Any chance you remember…      0
```

[6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 159571 entries, 0 to 159570
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   text    159571 non-null  object
 1   toxic   159571 non-null  int64
dtypes: int64(1), object(1)
memory usage: 2.4+ MB
```

[7]:
```
data.columns
```

[7]:
```
Index(['text', 'toxic'], dtype='object')
```

,

[8]:
```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
```

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

[8]:
```
True
```

[9]:
```
import string
punctuation = string.punctuation
```

```
[10]: noise = stopwords.words('english') + list(punctuation) + list('1234567890')
      def tokenizer(value):
          noise = stopwords.words('english') + list(punctuation) + list('1234567890')
          value = value.lower()
          a = word_tokenize(value)
          b = list()
          for el in a:
              if el not in noise:
                  if el.isdigit() == False:
                      if not el[0].isdigit():
                          b.append(el)
          stemmer = SnowballStemmer('english')
          stemmed_example = [stemmer.stem(w) for w in b]
          a = ' '.join(stemmed_example)
          return a
```

```
[11]: %%time
      data['text'] = data['text'].apply(tokenizer)
      data.head()
```

```
Wall time: 9min 35s
```

```
[11]:                                                      text  toxic
      0  explan edit made usernam hardcor metallica fan…      0
      1  d'aww match background colour 'm seem stuck th…      0
      2  hey man 'm realli tri edit war 's guy constant…      0
      3  `` ca n't make real suggest improv wonder sect…      0
      4                        sir hero chanc rememb page 's      0
```

, .

P.S. , 10

1.2

1.2.1 CountVectorizer Edition

```
[12]: data_cv = data.copy()
      x_train, x_val, y_train, y_val = train_test_split(data_cv.drop('toxic', axis=1),
                                                          data_cv['toxic'], test_size=0.
       ↪35, random_state=23)
```

, N 1 2.

```
[13]: count_vec = CountVectorizer(ngram_range=(1,2))
      x_train_count = count_vec.fit_transform(x_train['text'])
      x_train_count
```

```
[13]: <103721x1760255 sparse matrix of type '<class 'numpy.int64'>'
              with 6061886 stored elements in Compressed Sparse Row format>
```

```
[14]: print(x_train_count.shape)
      print(x_train.shape)
```

```
(103721, 1760255)
(103721, 1)
```

```
[15]: x_val_count = count_vec.transform(x_val['text'])
      x_val_count.shape
```

```
[15]: (55850, 1760255)
```

SVM SGD.

```
[16]: model_sgd1 = SGDClassifier(class_weight='balanced', random_state=131,␣
       ↪loss='hinge')
      model_sgd1.fit(x_train_count, y_train)
      pred1 = model_sgd1.predict(x_val_count)
      f1_score(y_val, pred1)
```

```
[16]: 0.7818863879957128
```

.

```
[17]: model_sgd2 = SGDClassifier(class_weight='balanced', random_state=131,␣
       ↪loss='log')
      model_sgd2.fit(x_train_count, y_train)
      pred2 = model_sgd2.predict(x_val_count)
      f1_score(y_val, pred2)
```

```
[17]: 0.7741500042183415
```

, SVM

SVM SGD

```
[18]: x_train, x_test, y_train, y_test = train_test_split(data_cv.drop('toxic',␣
       ↪axis=1),
                                                          data_cv['toxic'], test_size=0.
       ↪3, random_state=25433)
```

```
[20]: x_train_count = count_vec.fit_transform(x_train['text'])
      x_test_count = count_vec.transform(x_test['text'])
```

```
[21]: #model_sgd1 = SGDClassifier(class_weight='balanced', random_state=131,␣
       ↪loss='hinge', n_jobs=2)
```

```
#model_sgd1.fit(x_train, y_train)
model_sgd1.fit(x_train_count, y_train)
predt = model_sgd1.predict(x_test_count)
print('F1 Score on test data:',f1_score(y_test, predt))
```

```
F1 Score on test data: 0.7887890005288207
```

1.3

, SGD Classifier hinge loss. Countvectorizer N (1,1), , . , SVM

1.4 -

- ☒ Jupyter Notebook
- ☒
- ☒
- ☒
- ☒
- ☒ *F1* 0.75
- ☒

[ ]:

```

```