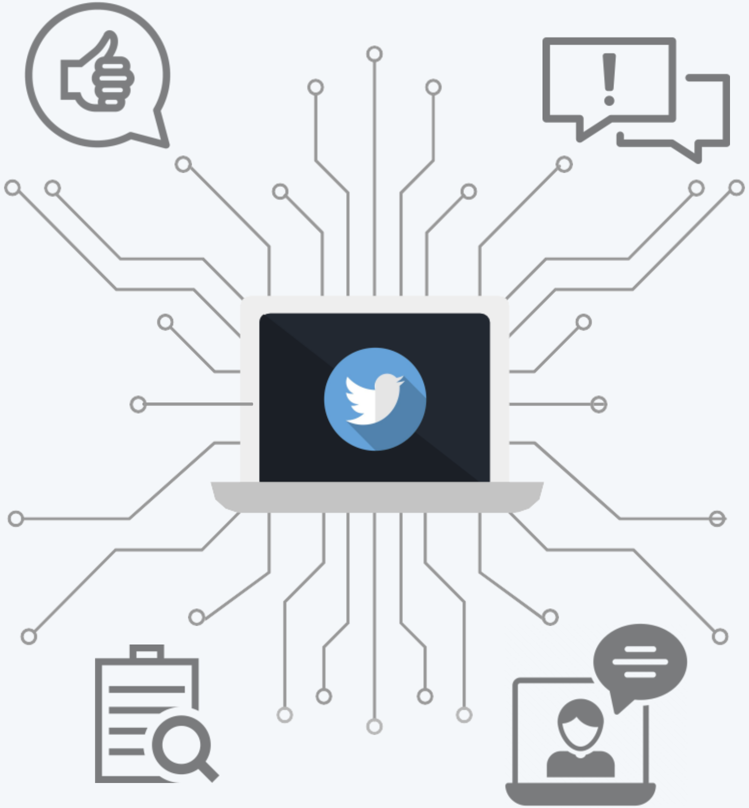
Twitter Hate Speech Detection

Can Content Moderation be Automated?

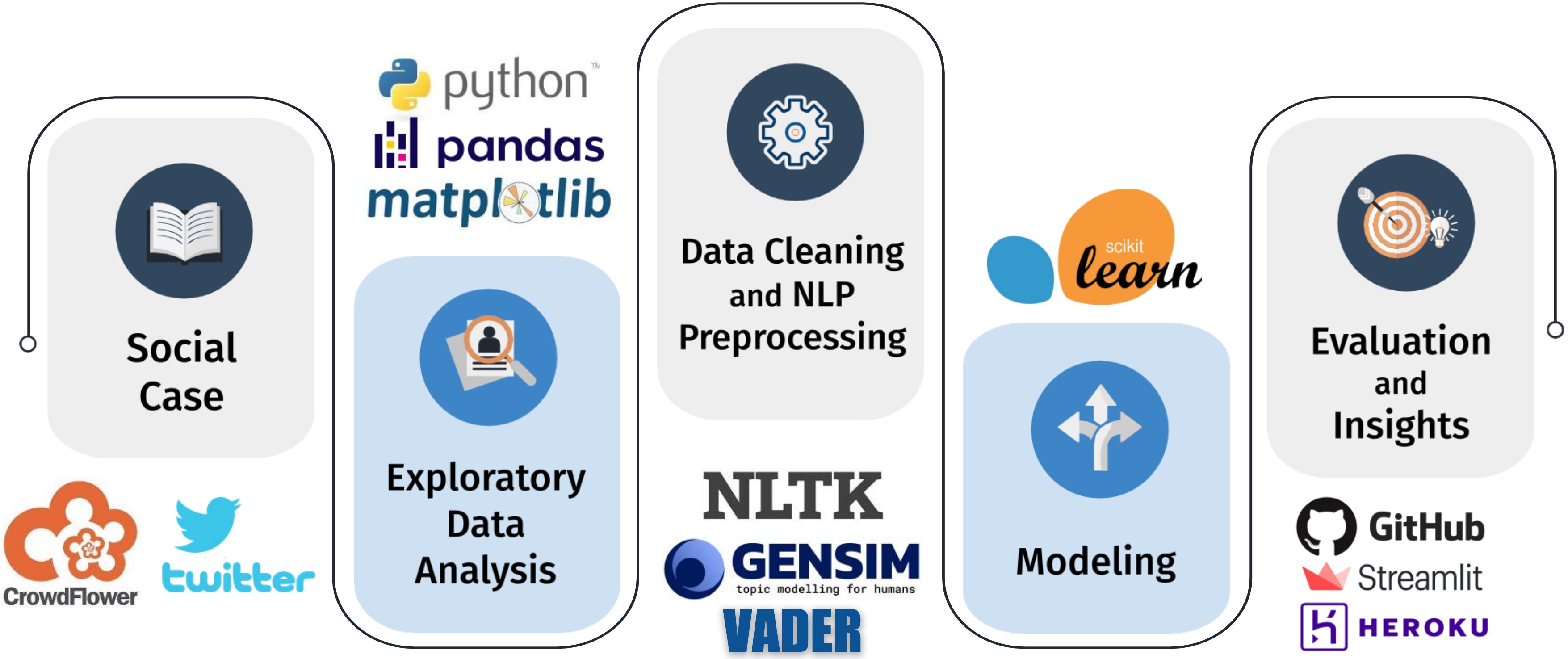
COINCENT DATA SCIENCE INTERN
KUKKAMALLA SANDESH

The ABTRACT

- **Every major tech company** uses third-party contractors
- **Automating** this process could **reduce labor exploitation**
- What is **Hate Speech?**



Objective-Process



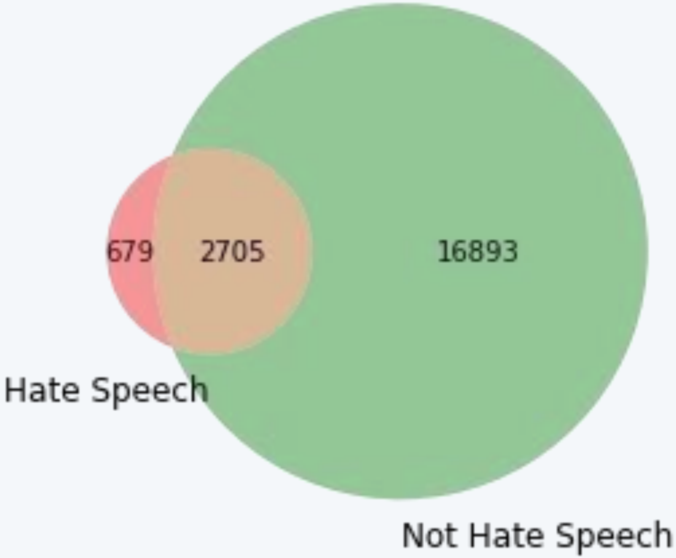
Social Case

Exploratory Data Analysis

Data Cleaning and NLP Preprocessing

Modeling

Evaluation and Insights



Unique Words per Label

Introduction

Sourced from 2017 Cornell University **research study.**

24,802 Tweets

20,277 Word Vocabulary

6% Hate Speech

94% Not Hate Speech

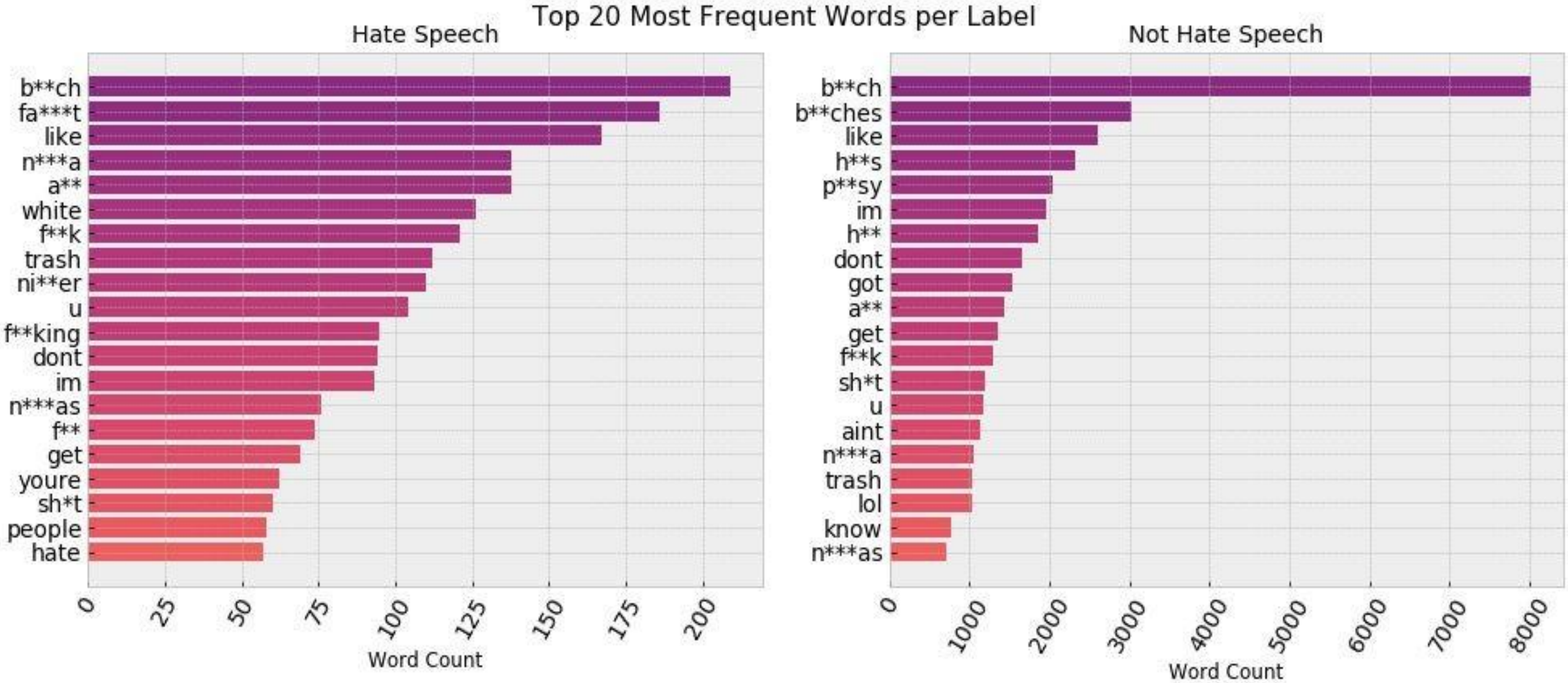
Data Analysis

1. What are the linguistic differences between hate speech and offensive language?

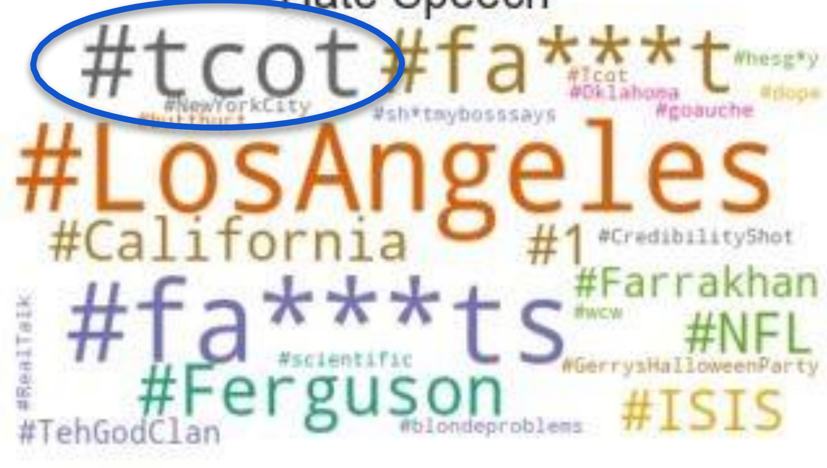
2. What are the popular hashtags of each tweet type?

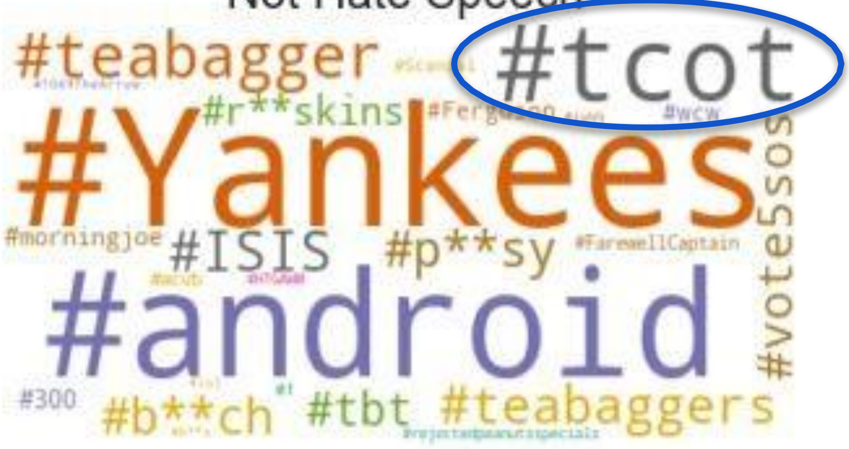
3. What is the overall polarity of the tweets?

What are the linguistic differences between hate speech and offensive language?

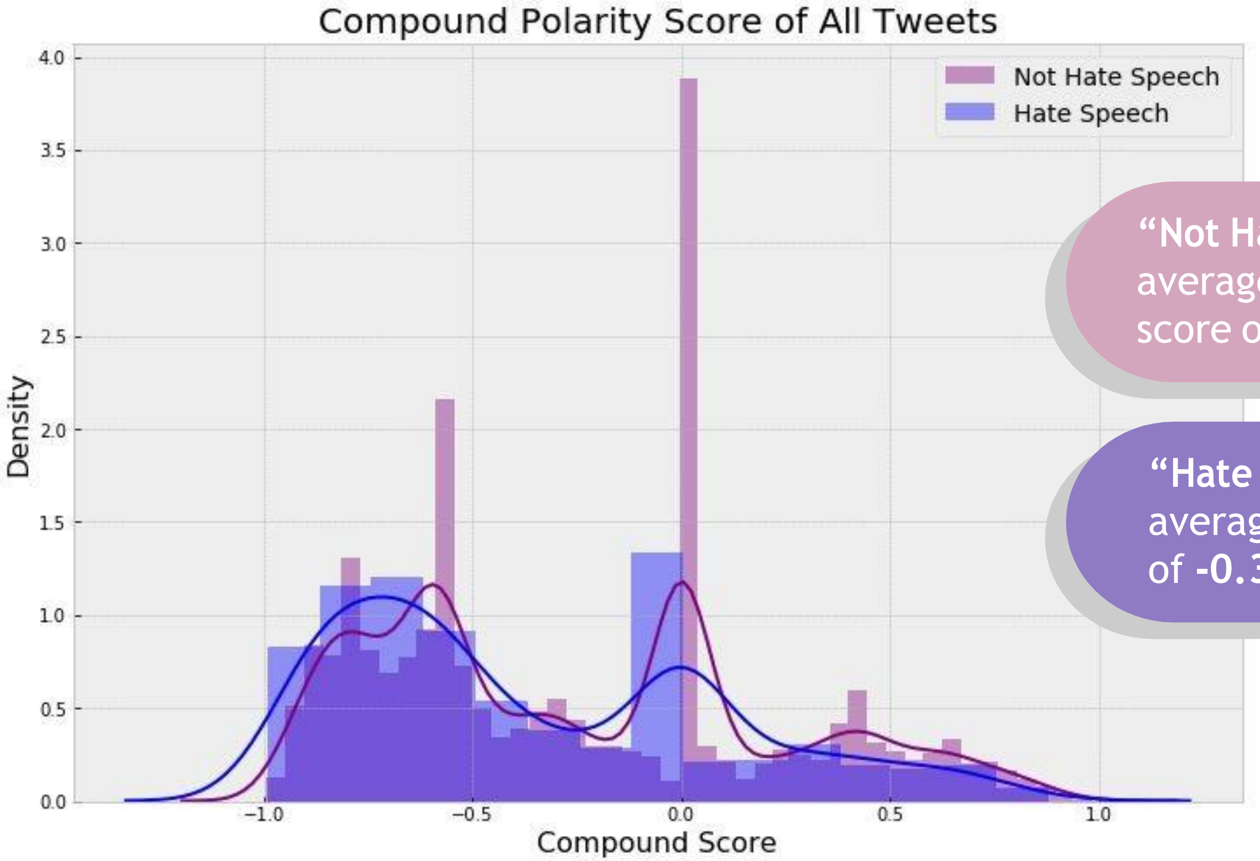


What are the **popular hashtags** of each tweet type?


Hate Speech


Not Hate Speech

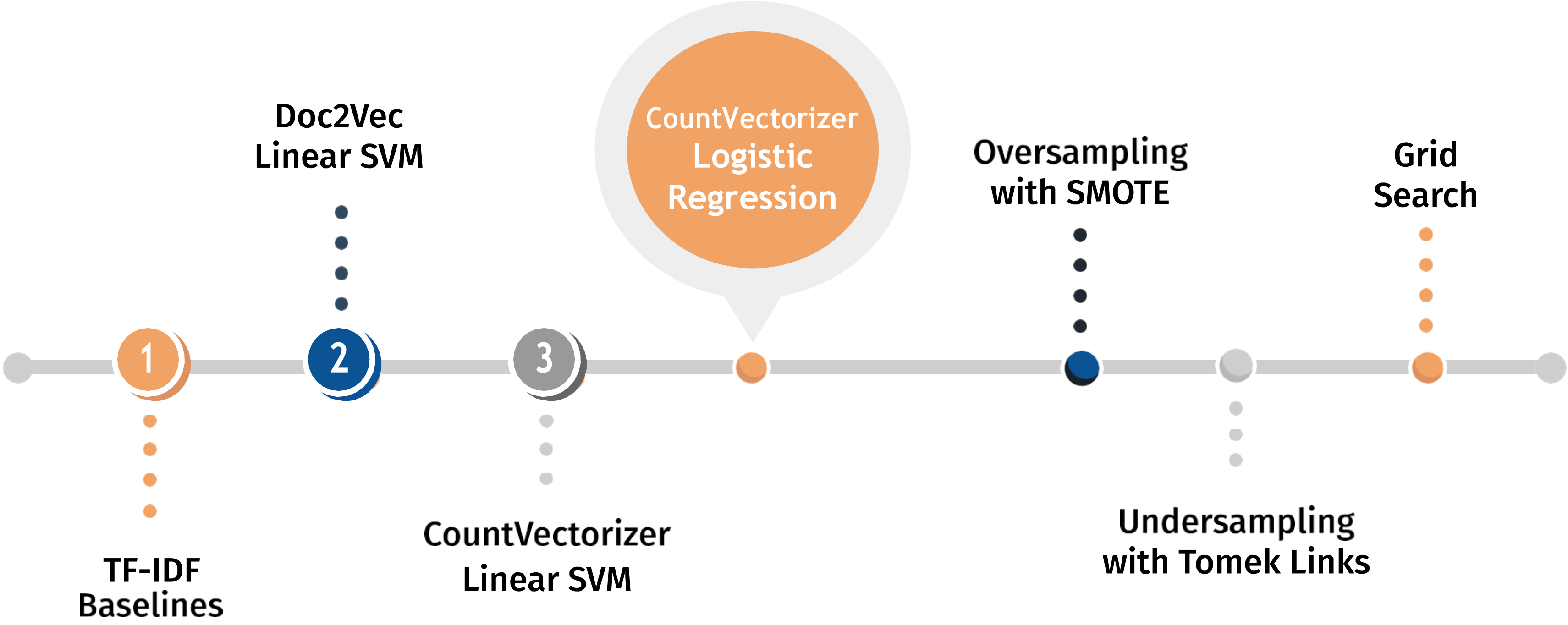
What is the **overall polarity** of the tweets?



"Not Hate Speech" tweets: average compound score of **-0.263**

"Hate Speech" tweets: average compound score of **-0.363**

Methodology



TF-IDF Baselines

Doc2Vec Linear SVM

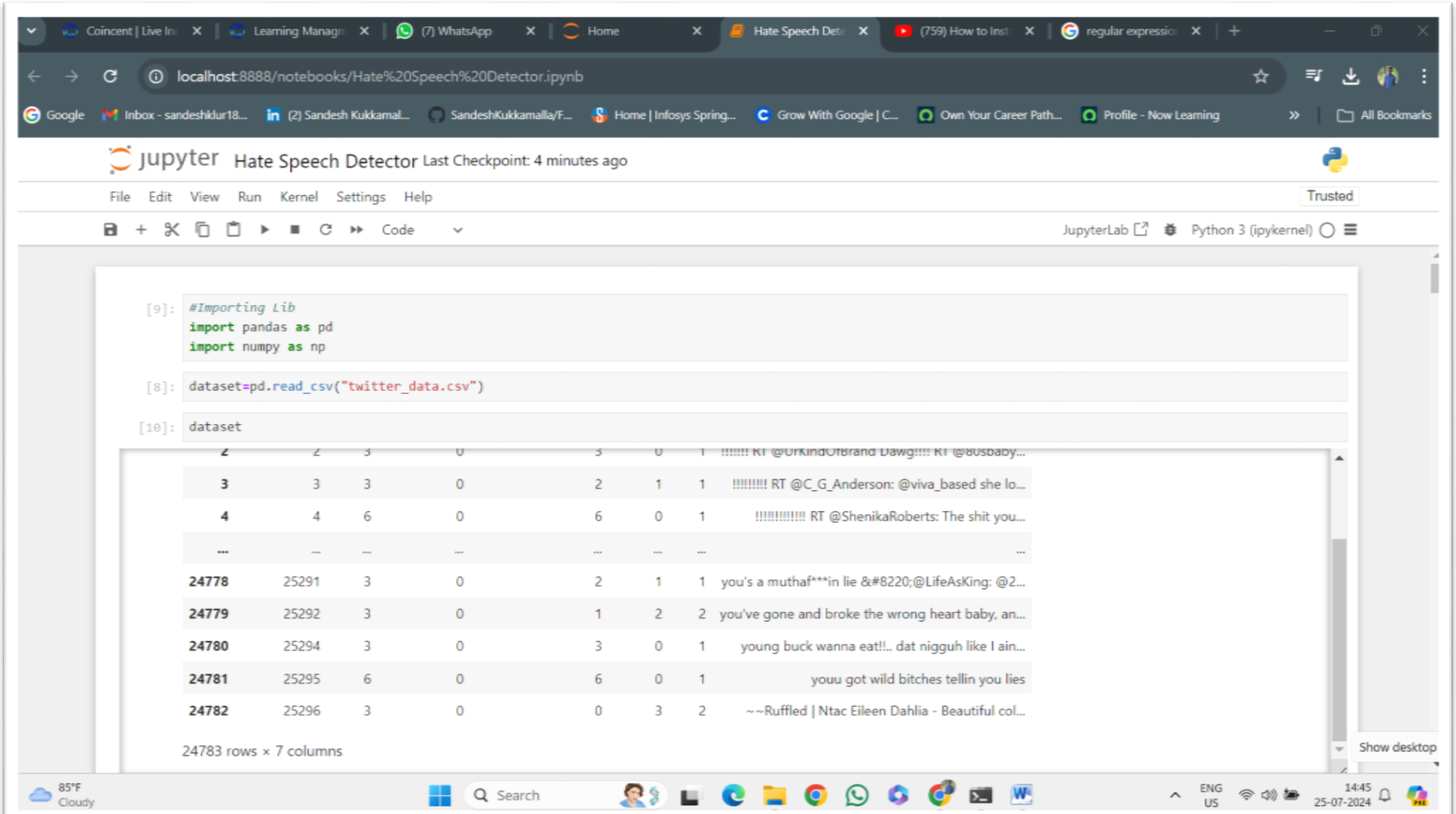
CountVectorizer Linear SVM

CountVectorizer Logistic Regression

Oversampling with SMOTE

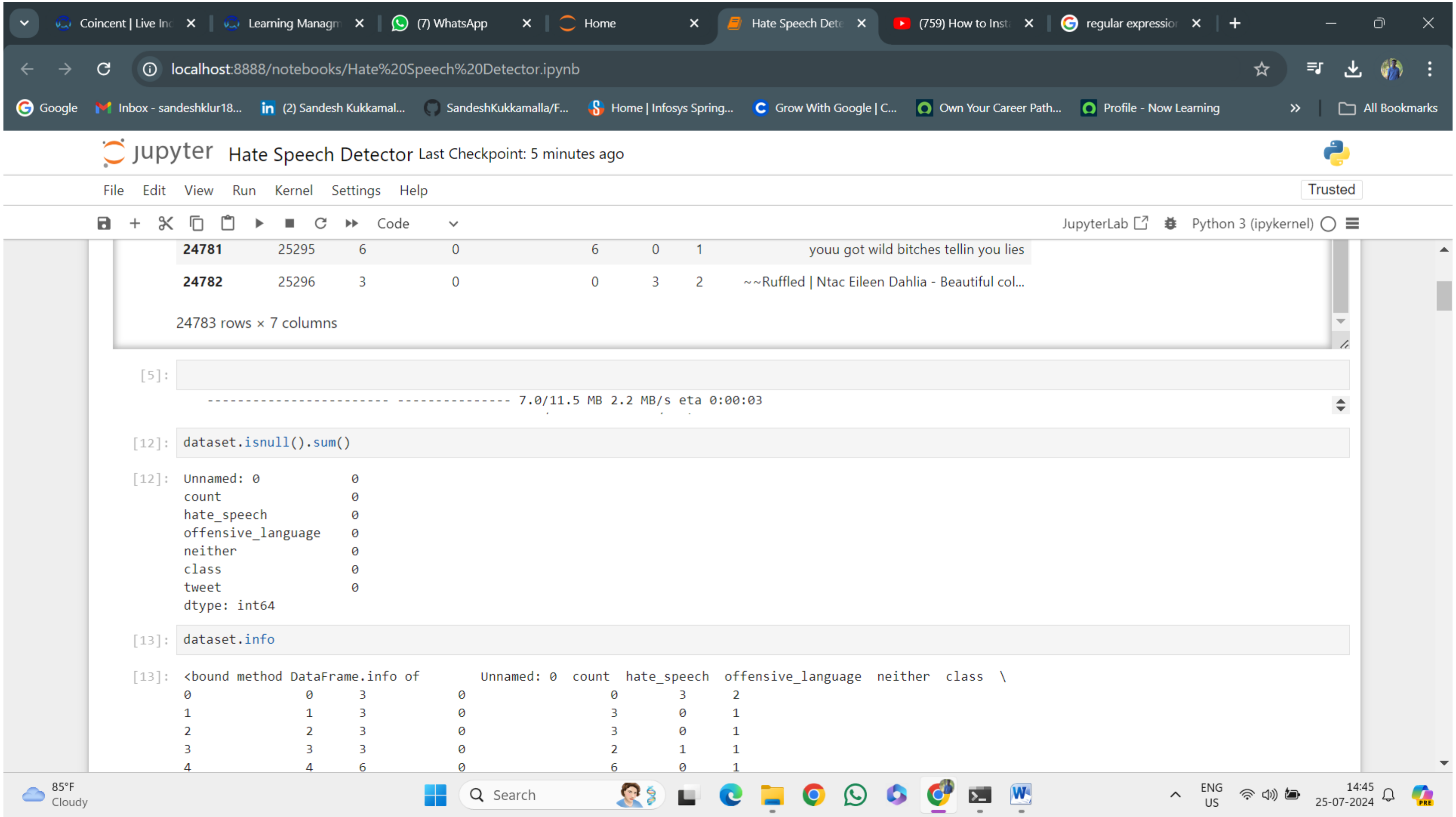
Undersampling with Tomek Links

Grid Search



```
[9]: #Importing Lib
     import pandas as pd
     import numpy as np

[8]: dataset=pd.read_csv("twitter_data.csv")

[10]: dataset
```

```
    2      2   3   0   3   0   1   !!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
    3      3   3   0   2   1   1   !!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
    4      4   6   0   6   0   1   !!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
  ...    ...  ... ... ... ... ...                                               ...
24778  25291   3   0   2   1   1   you's a muthaf***in lie &#8220;@LifeAsKing: @2...
24779  25292   3   0   1   2   2   you've gone and broke the wrong heart baby, an...
24780  25294   3   0   3   0   1   young buck wanna eat!!.. dat nigguh like I ain...
24781  25295   6   0   6   0   1   youu got wild bitches tellin you lies
24782  25296   3   0   0   3   2   ~~Ruffled | Ntac Eileen Dahlia - Beautiful col...

24783 rows × 7 columns
```



```
24781   25295   6   0   6   0   1        youu got wild bitches tellin you lies
24782   25296   3   0   0   3   2   ~~Ruffled | Ntac Eileen Dahlia - Beautiful col...

24783 rows × 7 columns
```

[5]:

```
    ------------------------ --------------- 7.0/11.5 MB 2.2 MB/s eta 0:00:03
```

[12]:

```
dataset.isnull().sum()
```

[12]:

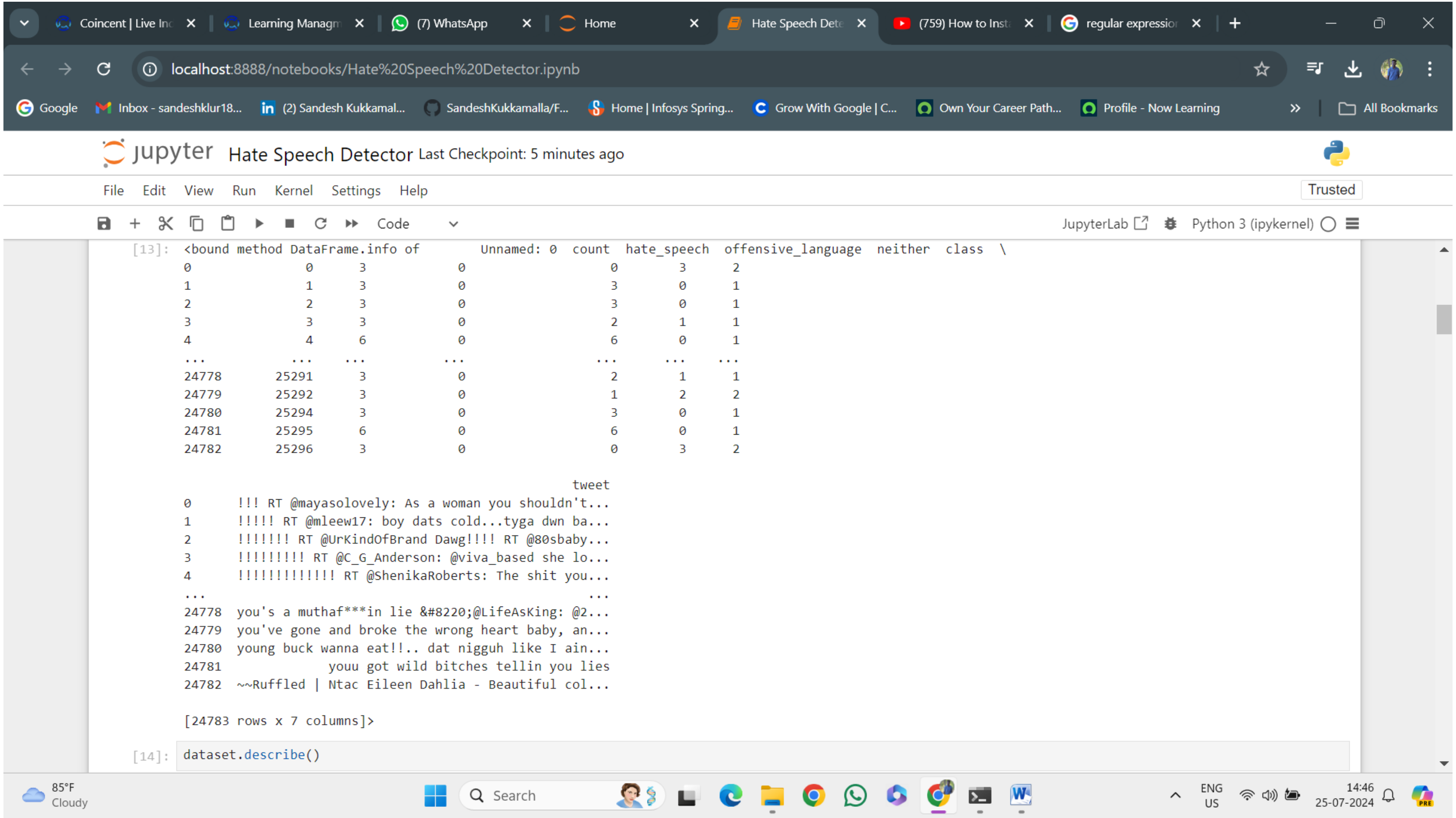
```
Unnamed: 0            0
count                 0
hate_speech           0
offensive_language    0
neither               0
class                 0
tweet                 0
dtype: int64
```

[13]:

```
dataset.info
```

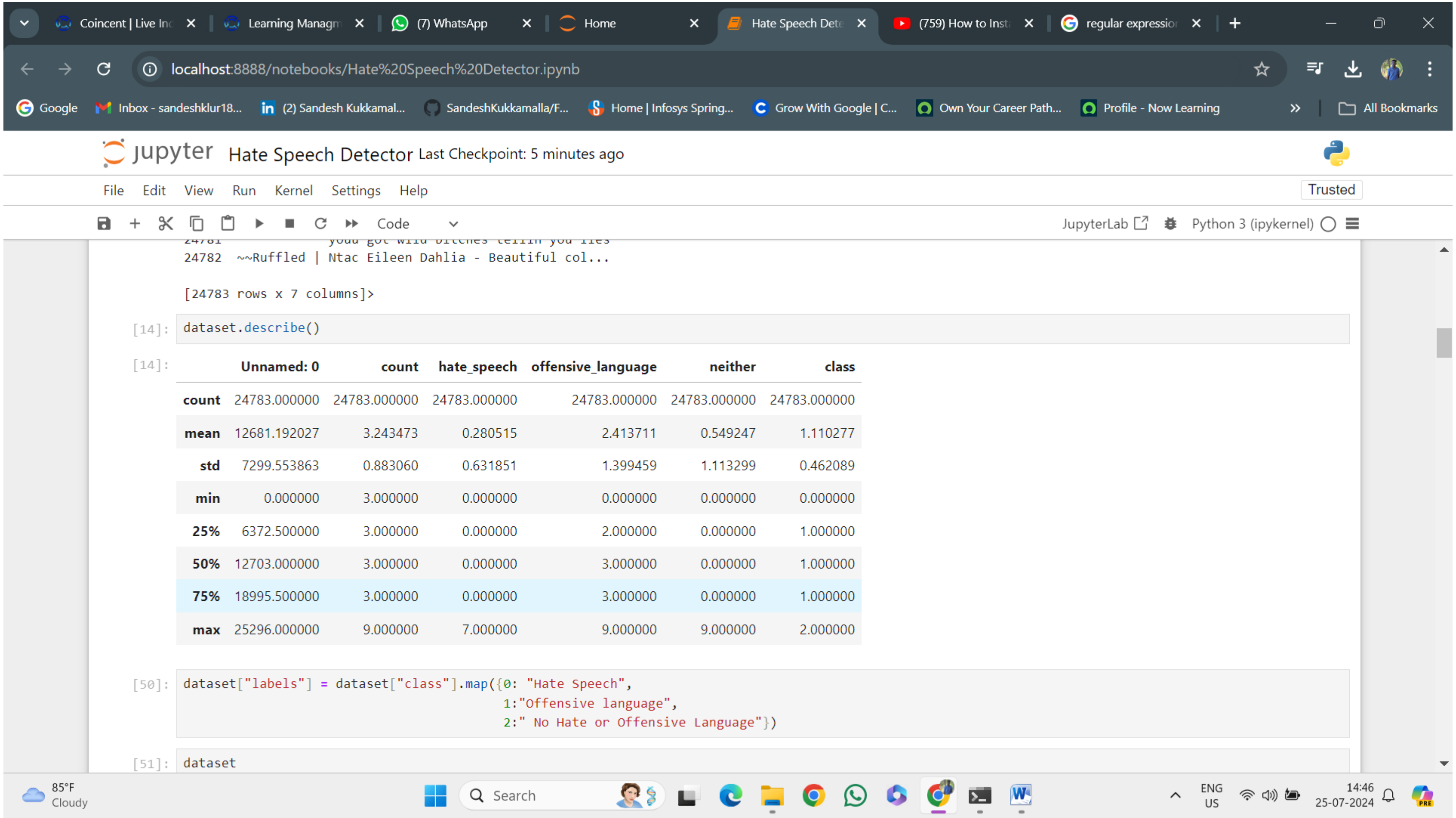
[13]:

```
<bound method DataFrame.info of        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
```



```
[13]: <bound method DataFrame.info of        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]>
```

```
[14]: dataset.describe()
```



```
24781              youd got wild bitches tellin you lies
24782  ~~Ruffled | Ntac Eileen Dahlia - Beautiful col...

[24783 rows x 7 columns]>
```

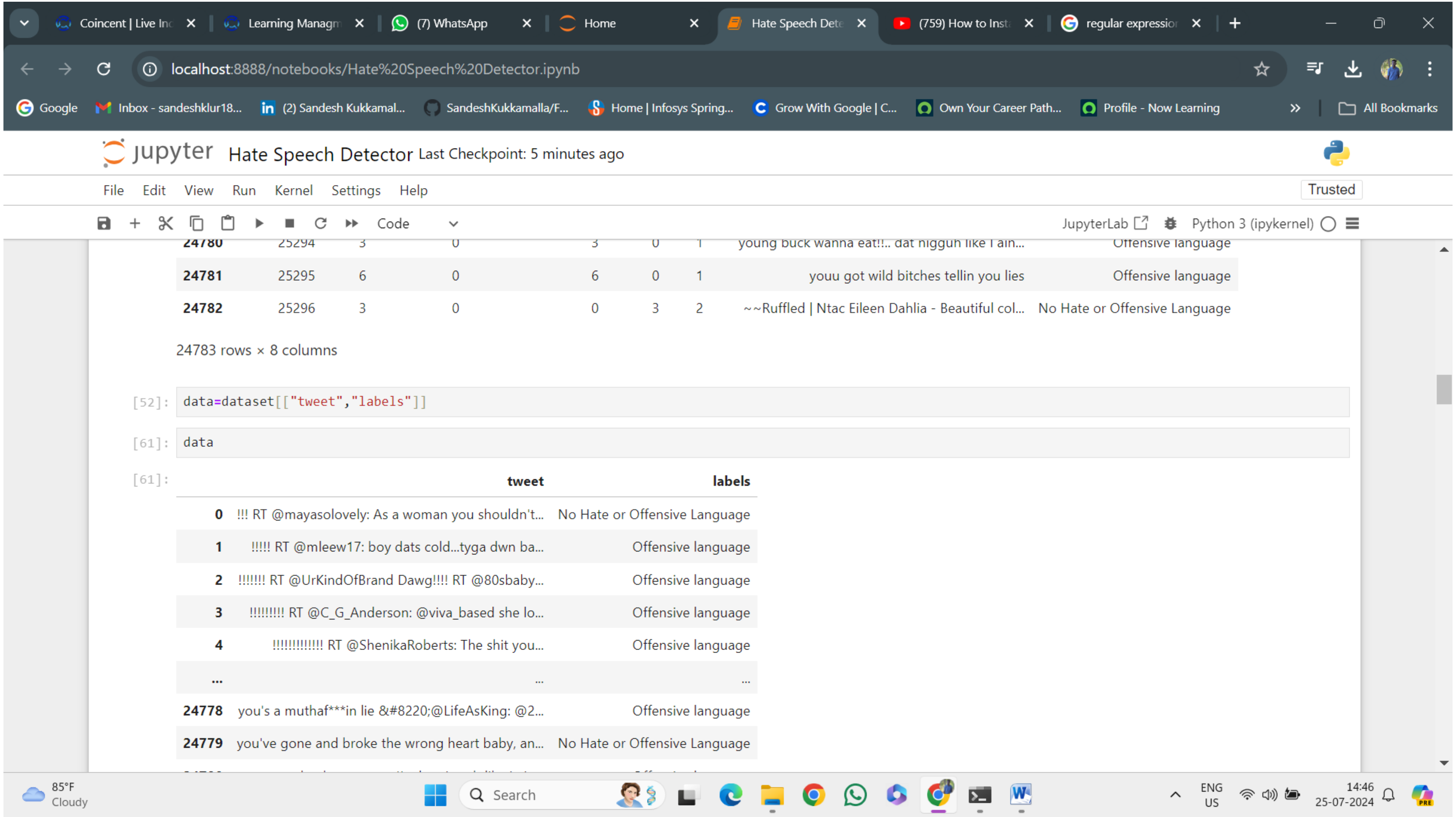
[14]:
```
dataset.describe()
```

[14]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class
count	24783.000000	24783.000000	24783.000000	24783.000000	24783.000000	24783.000000
mean	12681.192027	3.243473	0.280515	2.413711	0.549247	1.110277
std	7299.553863	0.883060	0.631851	1.399459	1.113299	0.462089
min	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000
25%	6372.500000	3.000000	0.000000	2.000000	0.000000	1.000000
50%	12703.000000	3.000000	0.000000	3.000000	0.000000	1.000000
75%	18995.500000	3.000000	0.000000	3.000000	0.000000	1.000000
max	25296.000000	9.000000	7.000000	9.000000	9.000000	2.000000

[50]:
```
dataset["labels"] = dataset["class"].map({0: "Hate Speech",
                                          1:"Offensive language",
                                          2:" No Hate or Offensive Language"})
```

[51]:
```
dataset
```



24780	25294	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...	Offensive language
24781	25295	6	0	6	0	1	youu got wild bitches tellin you lies	Offensive language
24782	25296	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	No Hate or Offensive Language

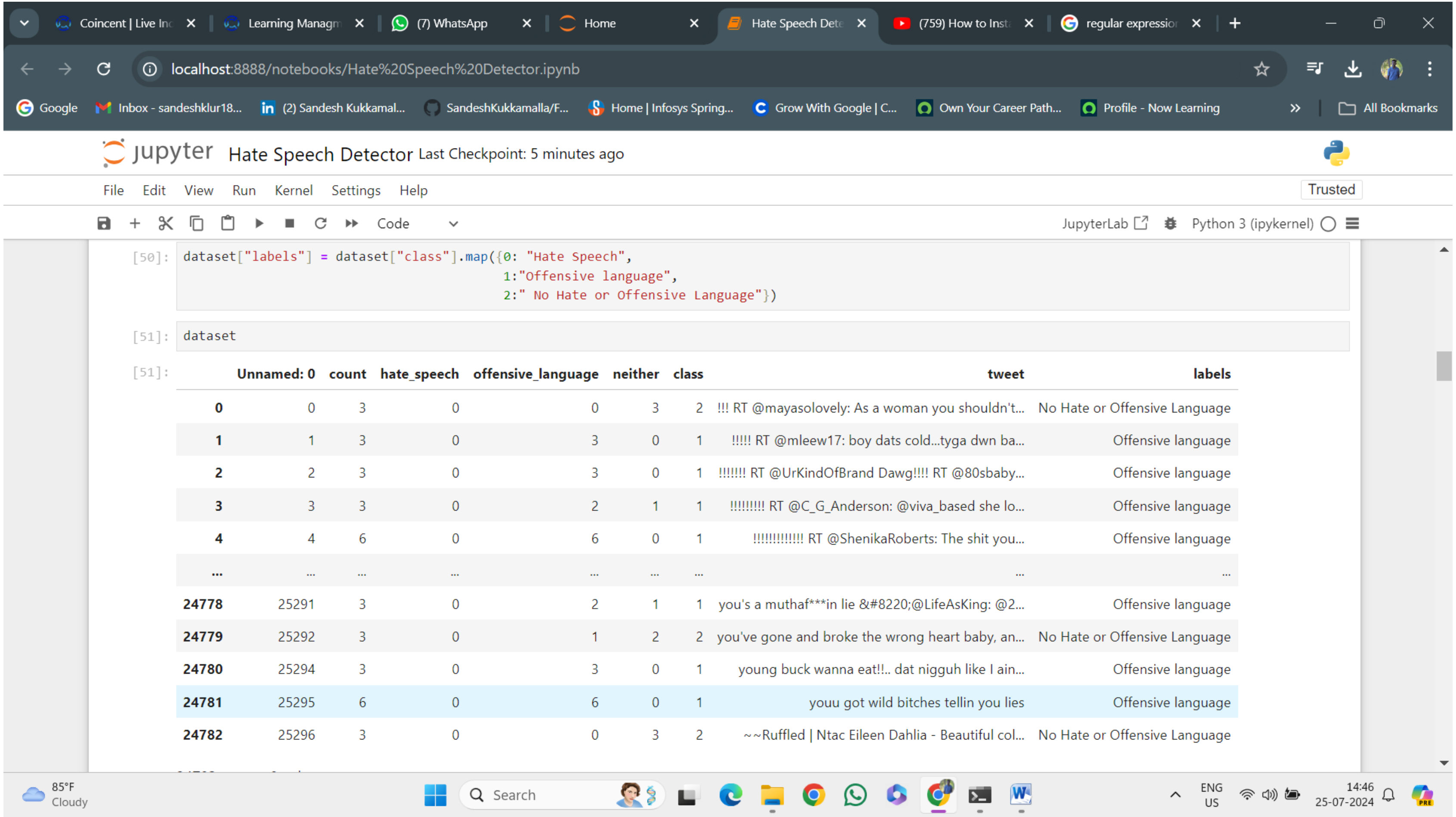
24783 rows × 8 columns

[52]:
```
data=dataset[["tweet","labels"]]
```

[61]:
```
data
```

[61]:

	tweet	labels
0	!!! RT @mayasolovely: As a woman you shouldn't...	No Hate or Offensive Language
1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	Offensive language
2	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	Offensive language
3	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	Offensive language
4	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	Offensive language
...	...	...
24778	you's a muthaf***in lie &#8220;@LifeAsKing: @2...	Offensive language
24779	you've gone and broke the wrong heart baby, an...	No Hate or Offensive Language

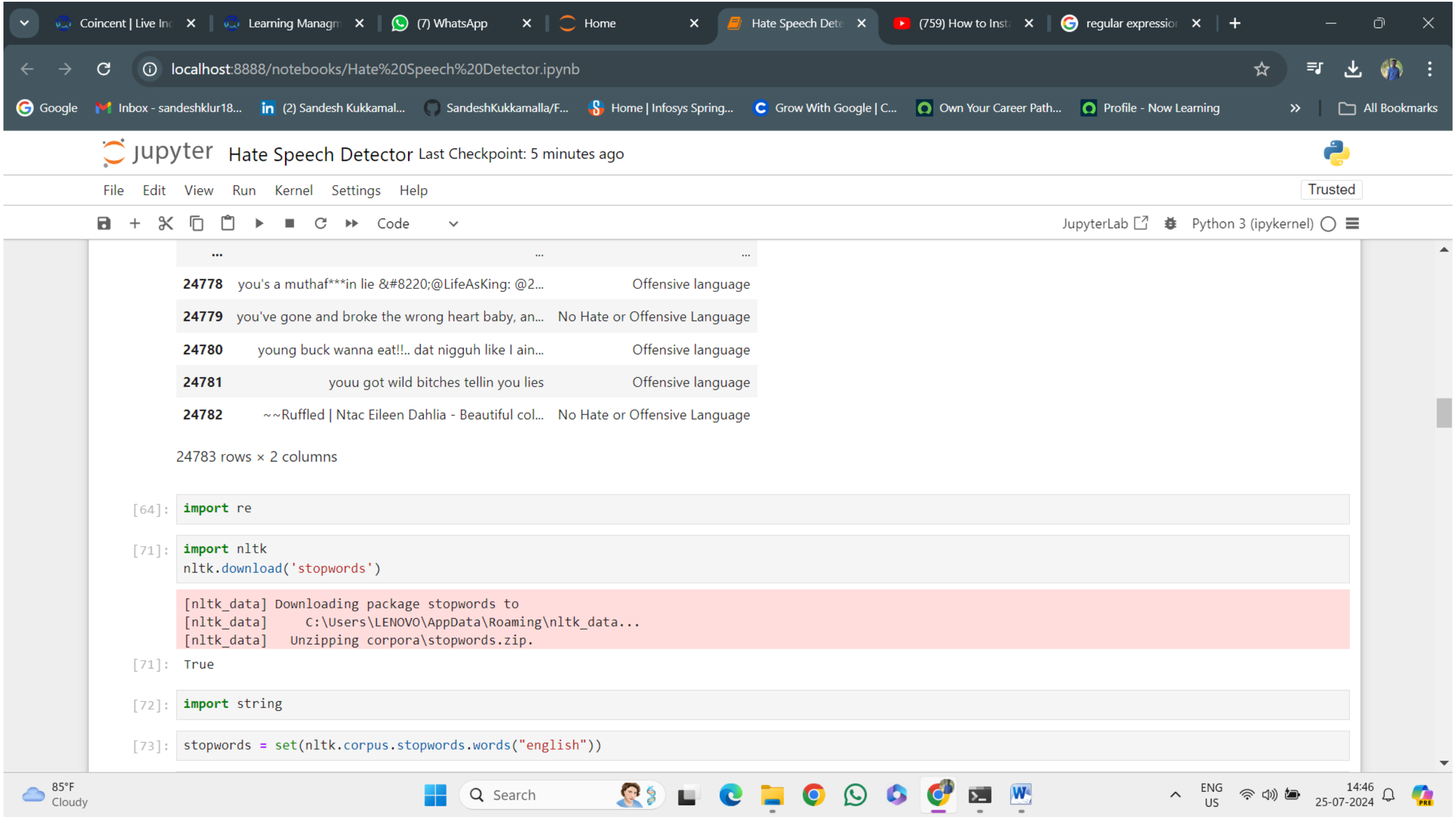


```
[50]: dataset["labels"] = dataset["class"].map({0: "Hate Speech",
                                               1:"Offensive language",
                                               2:" No Hate or Offensive Language"})
```

```
[51]: dataset
```

[51]:

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet	labels
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...	No Hate or Offensive Language
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...	Offensive language
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...	Offensive language
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...	Offensive language
4	4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...	Offensive language
...	...	...	...	...	...	...	...	...
24778	25291	3	0	2	1	1	you's a muthaf***in lie &#8220;@LifeAsKing: @2...	Offensive language
24779	25292	3	0	1	2	2	you've gone and broke the wrong heart baby, an...	No Hate or Offensive Language
24780	25294	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...	Offensive language
24781	25295	6	0	6	0	1	youu got wild bitches tellin you lies	Offensive language
24782	25296	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	No Hate or Offensive Language



...	...	...
24778	you's a muthaf***in lie &#8220;@LifeAsKing: @2...	Offensive language
24779	you've gone and broke the wrong heart baby, an...	No Hate or Offensive Language
24780	young buck wanna eat!!.. dat nigguh like I ain...	Offensive language
24781	youu got wild bitches tellin you lies	Offensive language
24782	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...	No Hate or Offensive Language

24783 rows × 2 columns

[64]:
```python
import re
```

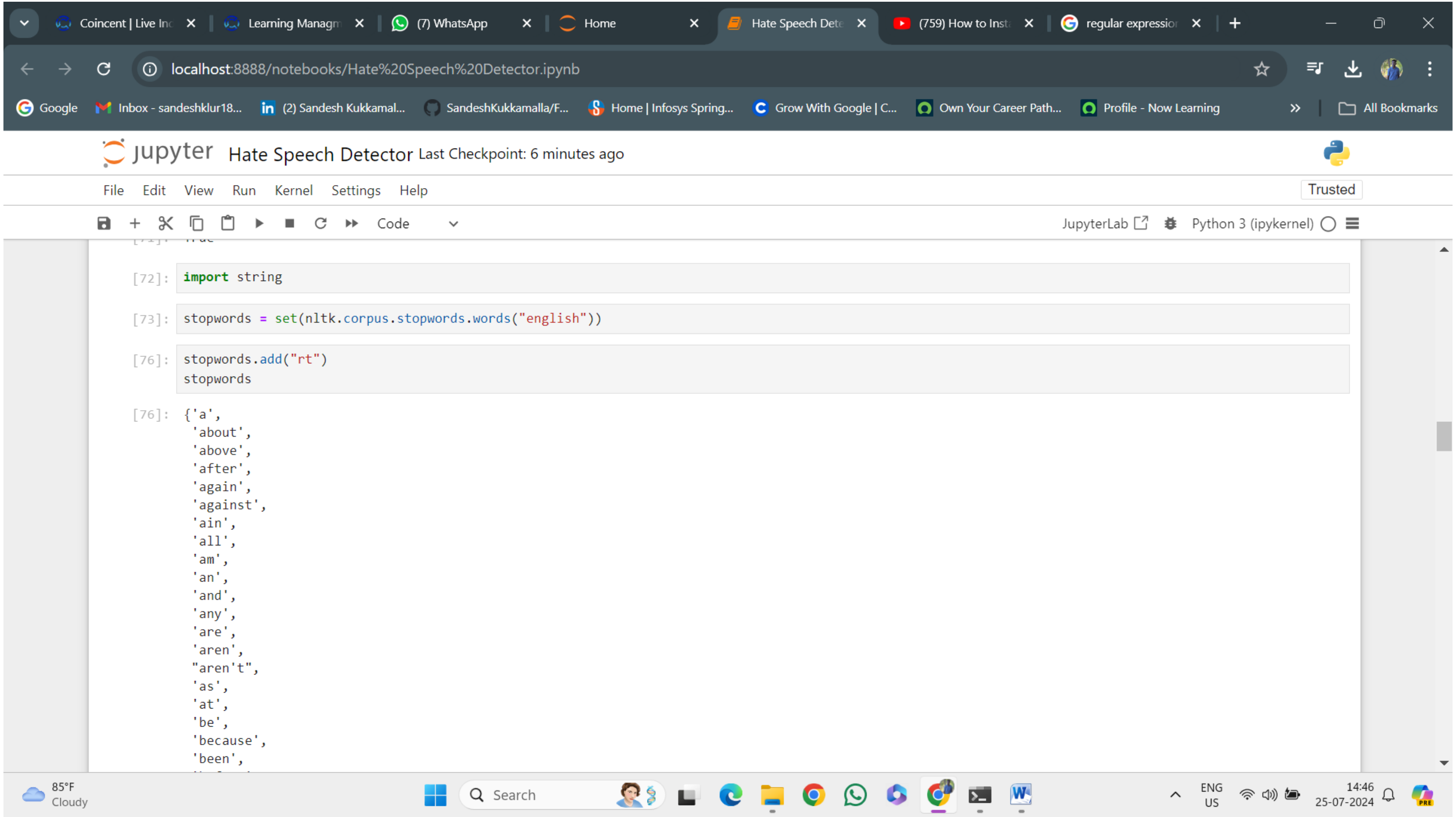
[71]:
```python
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\LENOVO\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

[71]:
```
True
```

[72]:
```python
import string
```

[73]:
```python
stopwords = set(nltk.corpus.stopwords.words("english"))
```

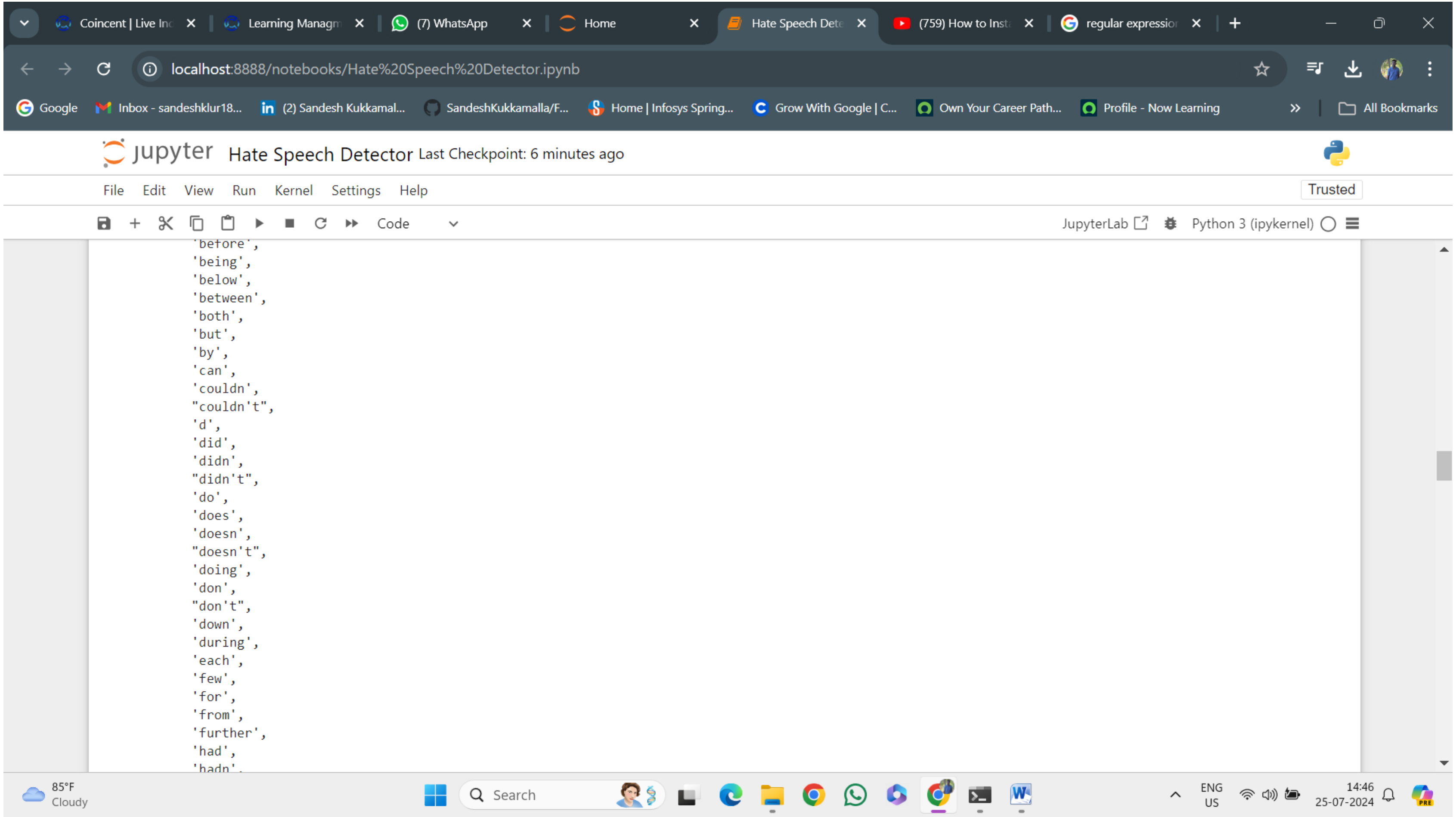


```
[72]: import string
```

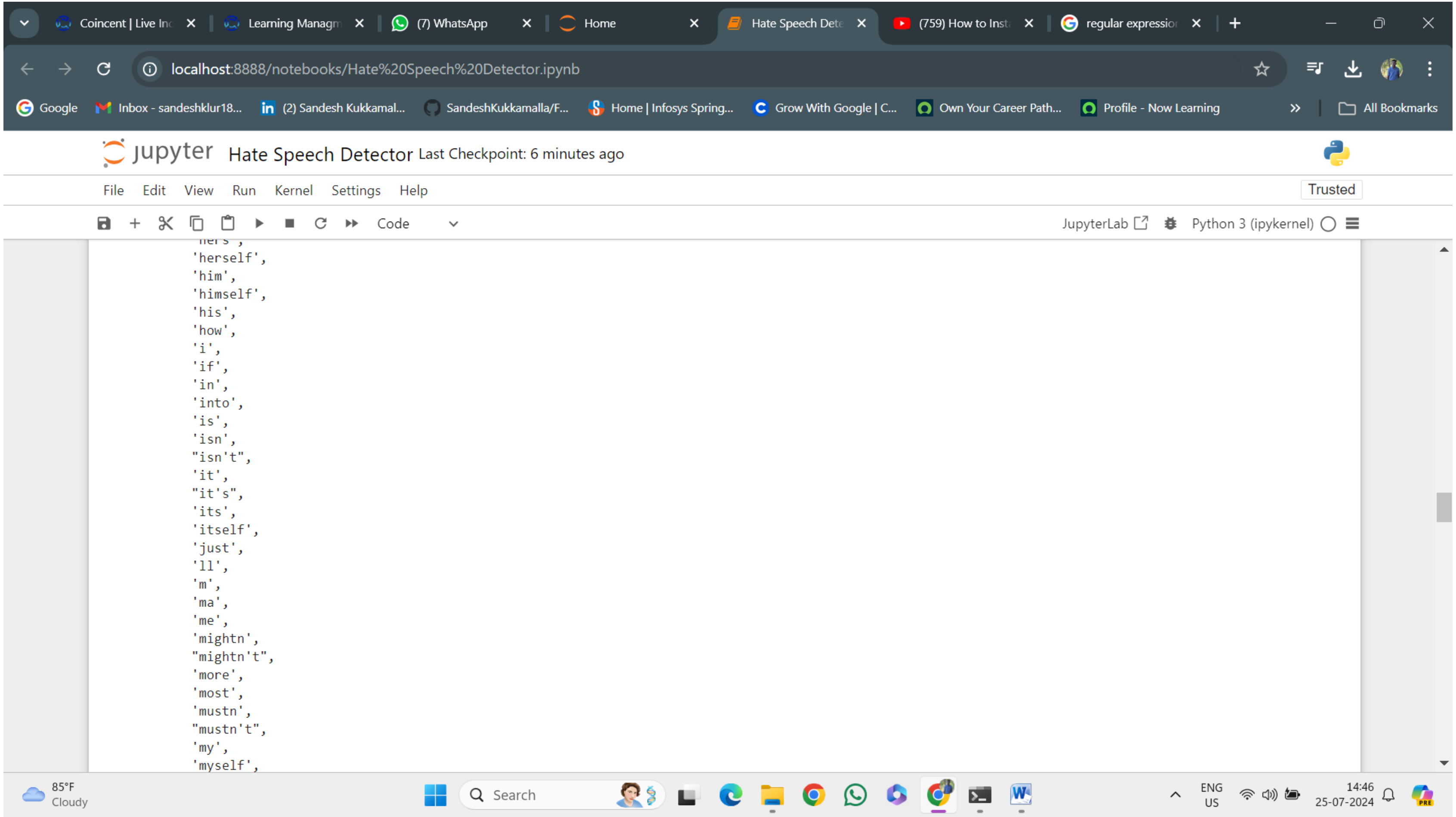
```
[73]: stopwords = set(nltk.corpus.stopwords.words("english"))
```

```
[76]: stopwords.add("rt")
      stopwords
```

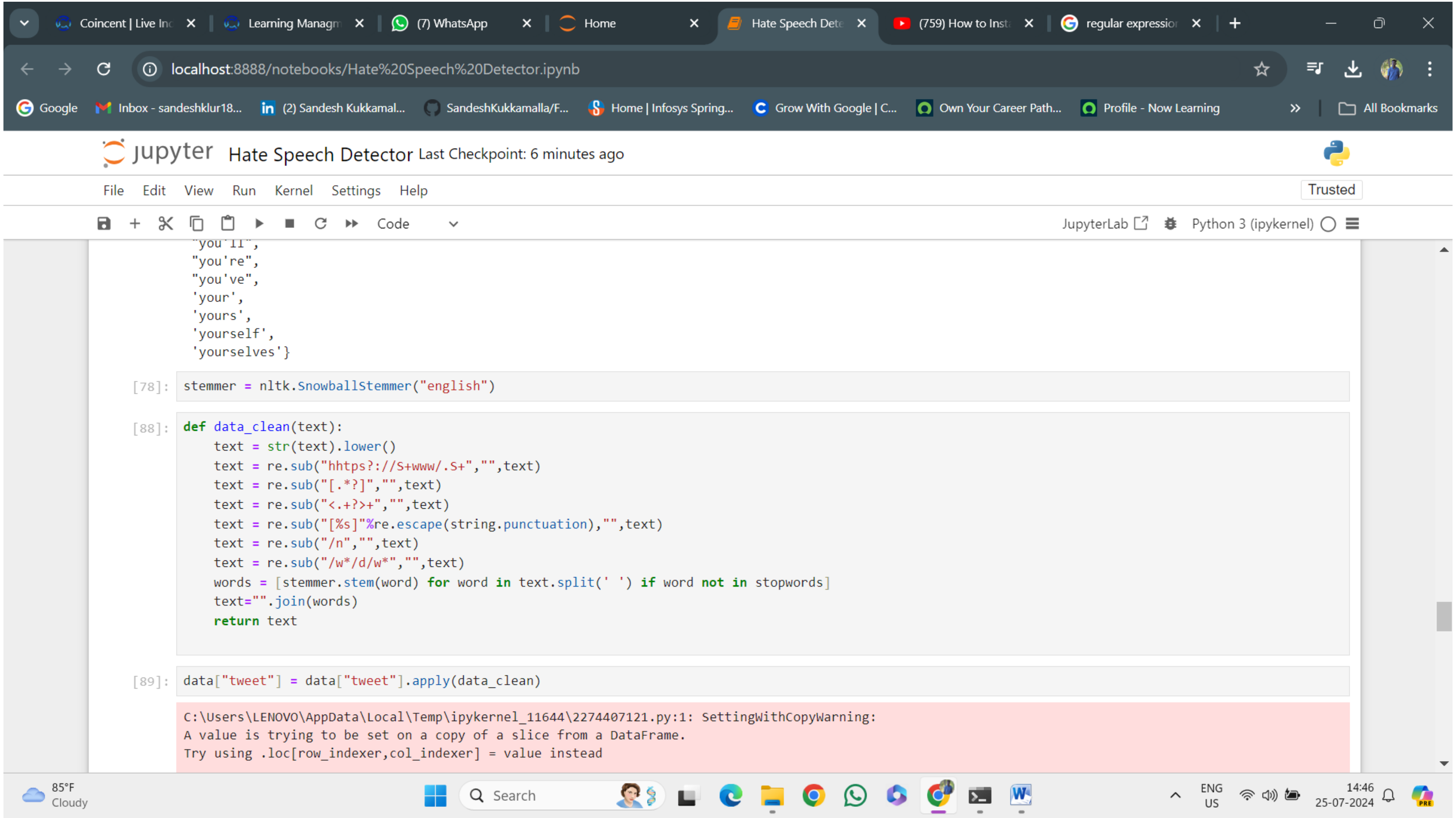
```
[76]: {'a',
       'about',
       'above',
       'after',
       'again',
       'against',
       'ain',
       'all',
       'am',
       'an',
       'and',
       'any',
       'are',
       'aren',
       "aren't",
       'as',
       'at',
       'be',
       'because',
       'been',
```



```
 'before',
 'being',
 'below',
 'between',
 'both',
 'but',
 'by',
 'can',
 'couldn',
 "couldn't",
 'd',
 'did',
 'didn',
 "didn't",
 'do',
 'does',
 'doesn',
 "doesn't",
 'doing',
 'don',
 "don't",
 'down',
 'during',
 'each',
 'few',
 'for',
 'from',
 'further',
 'had',
 'hadn',
```



```
 'hers',
 'herself',
 'him',
 'himself',
 'his',
 'how',
 'i',
 'if',
 'in',
 'into',
 'is',
 'isn',
 "isn't",
 'it',
 "it's",
 'its',
 'itself',
 'just',
 'll',
 'm',
 'ma',
 'me',
 'mightn',
 "mightn't",
 'more',
 'most',
 'mustn',
 "mustn't",
 'my',
 'myself',
```



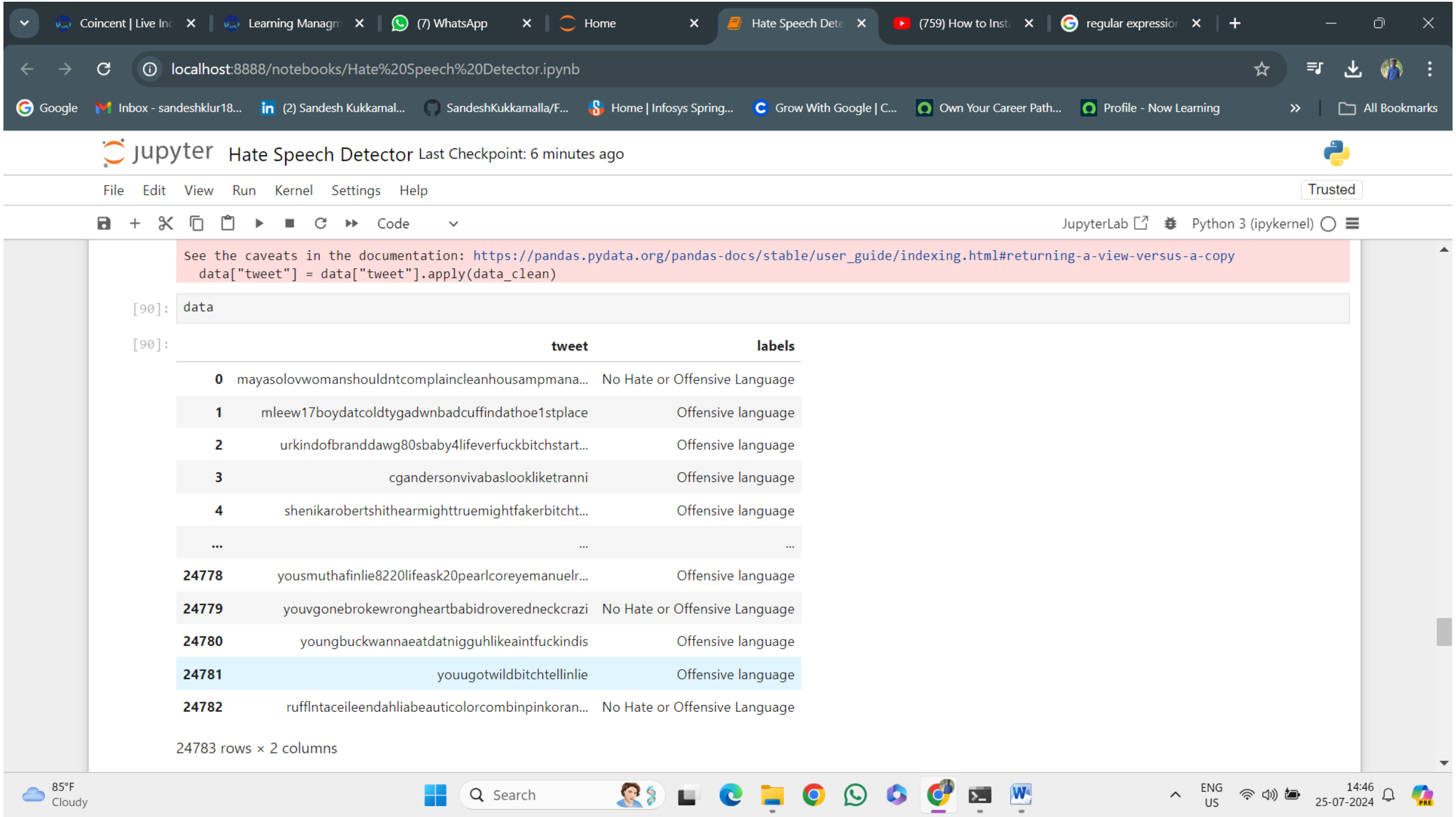
```
 "you'll",
 "you're",
 "you've",
 'your',
 'yours',
 'yourself',
 'yourselves'}
```

```
[78]: stemmer = nltk.SnowballStemmer("english")
```

```
[88]: def data_clean(text):
          text = str(text).lower()
          text = re.sub("hhttps?://S+www/.S+","",text)
          text = re.sub("[.*?]","",text)
          text = re.sub("<.+?>+","",text)
          text = re.sub("[%s]"%re.escape(string.punctuation),"",text)
          text = re.sub("/n","",text)
          text = re.sub("/w*/d/w*","",text)
          words = [stemmer.stem(word) for word in text.split(' ') if word not in stopwords]
          text="".join(words)
          return text
```

```
[89]: data["tweet"] = data["tweet"].apply(data_clean)
```

```
C:\Users\LENOVO\AppData\Local\Temp\ipykernel_11644\2274407121.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```



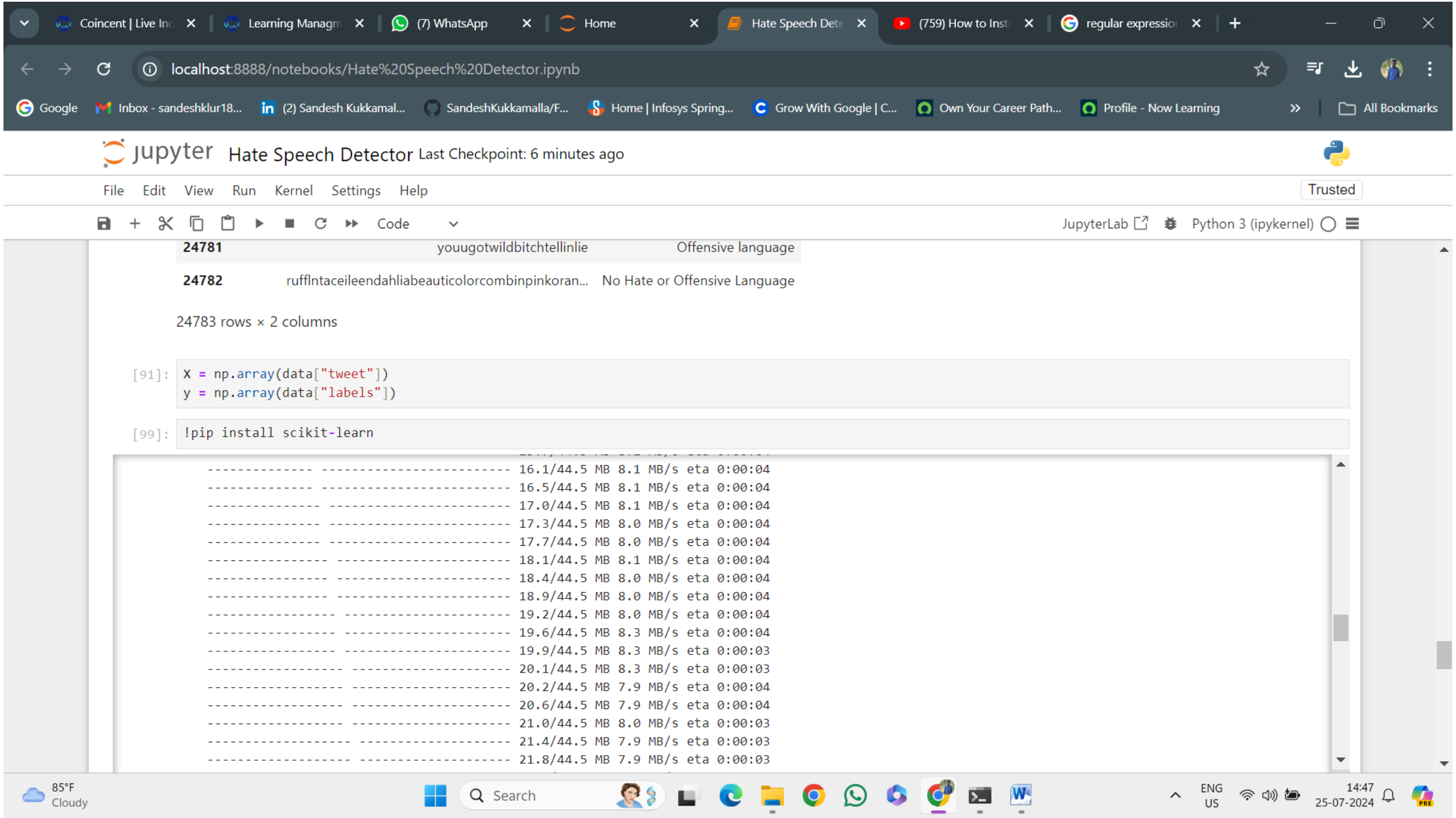
```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data["tweet"] = data["tweet"].apply(data_clean)
```

```
[90]: data
```

[90]:

	tweet	labels
0	mayasolovwomanshouldntcomplaincleanhousampmana...	No Hate or Offensive Language
1	mleew17boydatcoldtygadwnbadcuffindathoe1stplace	Offensive language
2	urkindofbranddawg80sbaby4lifeverfuckbitchstart...	Offensive language
3	cgandersonvivabaslookliketranni	Offensive language
4	shenikarobertshithearmighttruemightfakerbitcht...	Offensive language
...	...	...
24778	yousmuthafinlie8220lifeask20pearlcoreyemanuelr...	Offensive language
24779	youvgonebrokewrongheartbabidroveredneckcrazi	No Hate or Offensive Language
24780	youngbuckwannaeatdatnigguhlikeaintfuckindis	Offensive language
24781	youugotwildbitchtellinlie	Offensive language
24782	ruffIntaceileendahliabeauticolorcombinpinkoran...	No Hate or Offensive Language

24783 rows × 2 columns



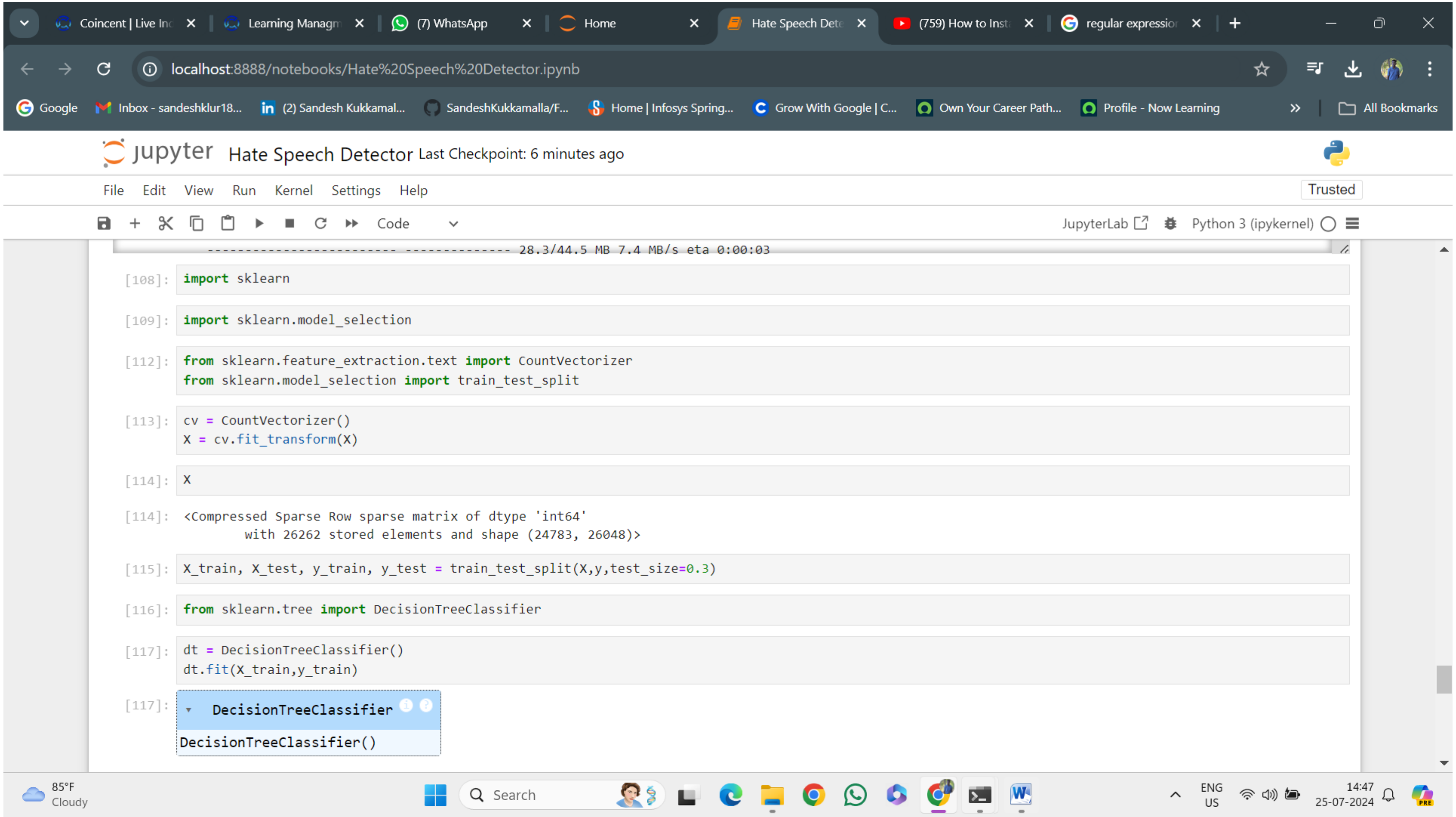
24781	youugotwildbitchtellinlie	Offensive language
24782	rufflntaceileendahliabeauticolorcombinpinkoran...	No Hate or Offensive Language

24783 rows × 2 columns

```
[91]: X = np.array(data["tweet"])
      y = np.array(data["labels"])
```

```
[99]: !pip install scikit-learn
```

```
    -------------- ------------------------- 16.1/44.5 MB 8.1 MB/s eta 0:00:04
    -------------- ------------------------- 16.5/44.5 MB 8.1 MB/s eta 0:00:04
    --------------- ------------------------ 17.0/44.5 MB 8.1 MB/s eta 0:00:04
    --------------- ------------------------ 17.3/44.5 MB 8.0 MB/s eta 0:00:04
    --------------- ------------------------ 17.7/44.5 MB 8.0 MB/s eta 0:00:04
    ---------------- ----------------------- 18.1/44.5 MB 8.1 MB/s eta 0:00:04
    ---------------- ----------------------- 18.4/44.5 MB 8.0 MB/s eta 0:00:04
    ---------------- ----------------------- 18.9/44.5 MB 8.0 MB/s eta 0:00:04
    ----------------- ---------------------- 19.2/44.5 MB 8.0 MB/s eta 0:00:04
    ----------------- ---------------------- 19.6/44.5 MB 8.3 MB/s eta 0:00:04
    ----------------- ---------------------- 19.9/44.5 MB 8.3 MB/s eta 0:00:03
    ------------------ --------------------- 20.1/44.5 MB 8.3 MB/s eta 0:00:03
    ------------------ --------------------- 20.2/44.5 MB 7.9 MB/s eta 0:00:04
    ------------------ --------------------- 20.6/44.5 MB 7.9 MB/s eta 0:00:04
    ------------------ --------------------- 21.0/44.5 MB 8.0 MB/s eta 0:00:03
    ------------------- -------------------- 21.4/44.5 MB 7.9 MB/s eta 0:00:03
    ------------------- -------------------- 21.8/44.5 MB 7.9 MB/s eta 0:00:03
```



```
   ------------------------ -------------- 28.3/44.5 MB 7.4 MB/s eta 0:00:03
```

```
[108]: import sklearn
```

```
[109]: import sklearn.model_selection
```

```
[112]: from sklearn.feature_extraction.text import CountVectorizer
       from sklearn.model_selection import train_test_split
```

```
[113]: cv = CountVectorizer()
       X = cv.fit_transform(X)
```

```
[114]: X
```

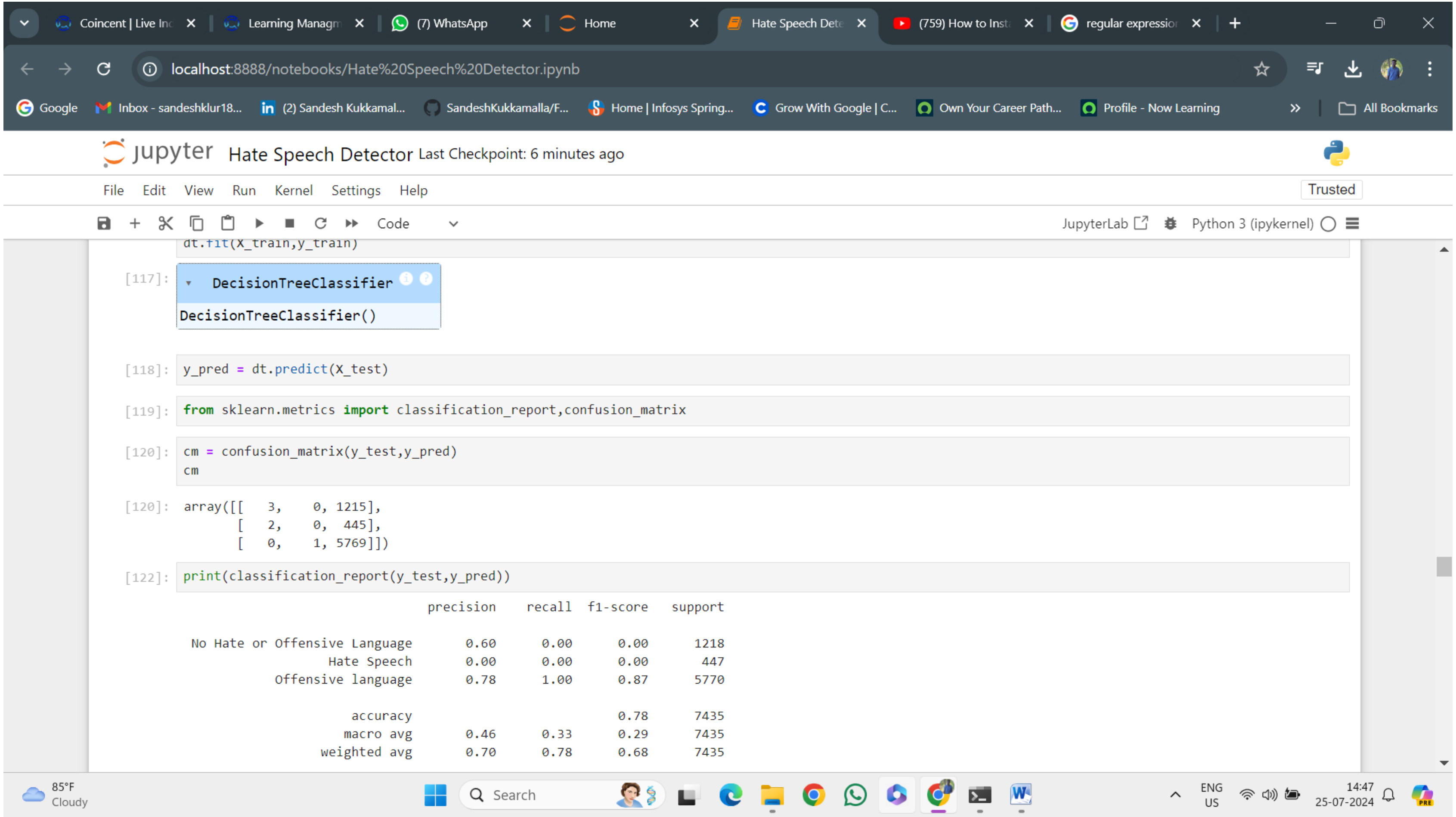
```
[114]: <Compressed Sparse Row sparse matrix of dtype 'int64'
               with 26262 stored elements and shape (24783, 26048)>
```

```
[115]: X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3)
```

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

```
[117]: dt = DecisionTreeClassifier()
       dt.fit(X_train,y_train)
```

```
[117]: ▾ DecisionTreeClassifier
        DecisionTreeClassifier()
```



```
dt.fit(X_train,y_train)
```

```
[117]:  ▾ DecisionTreeClassifier
        DecisionTreeClassifier()
```

```
[118]:  y_pred = dt.predict(X_test)
```

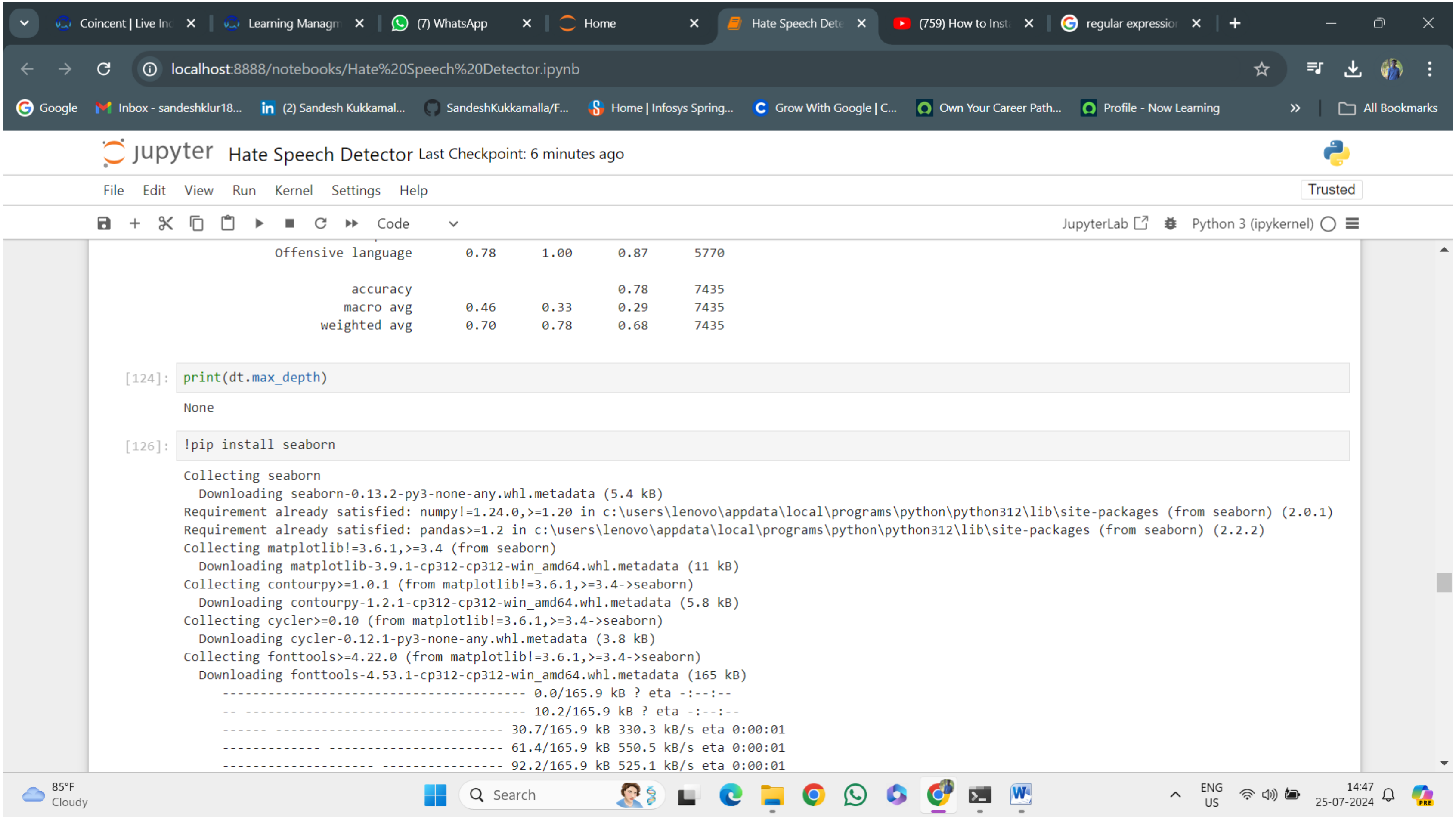
```
[119]:  from sklearn.metrics import classification_report,confusion_matrix
```

```
[120]:  cm = confusion_matrix(y_test,y_pred)
        cm
```

```
[120]:  array([[   3,    0, 1215],
               [   2,    0,  445],
               [   0,    1, 5769]])
```

```
[122]:  print(classification_report(y_test,y_pred))
```

```
                               precision    recall  f1-score   support

No Hate or Offensive Language       0.60      0.00      0.00      1218
                  Hate Speech       0.00      0.00      0.00       447
           Offensive language       0.78      1.00      0.87      5770

                     accuracy                           0.78      7435
                    macro avg       0.46      0.33      0.29      7435
                 weighted avg       0.70      0.78      0.68      7435
```



```
       Offensive language       0.78      1.00      0.87      5770

                 accuracy                           0.78      7435
                macro avg       0.46      0.33      0.29      7435
             weighted avg       0.70      0.78      0.68      7435
```

[124]:

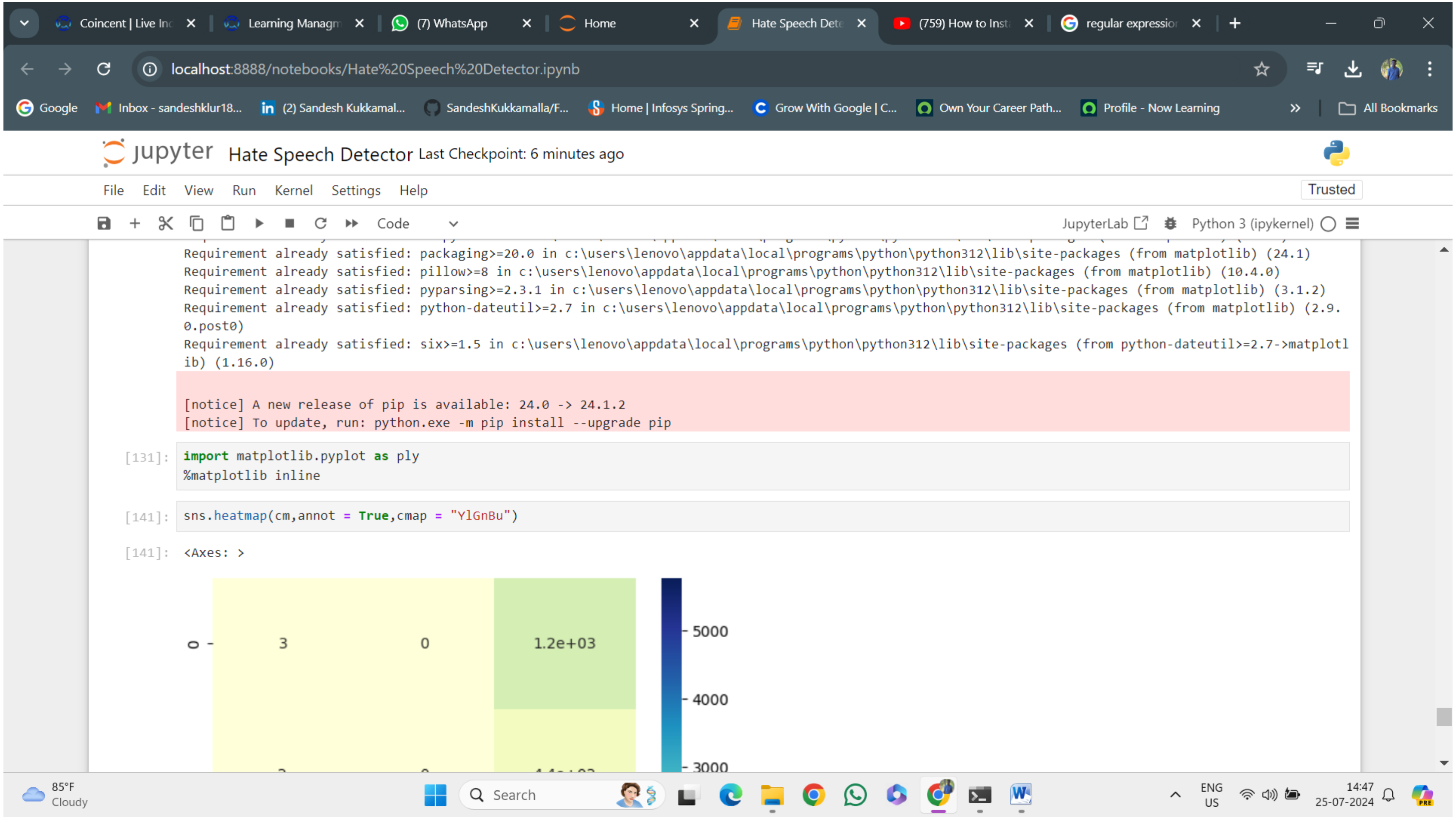
```
print(dt.max_depth)
```

```
None
```

[126]:

```
!pip install seaborn
```

```
Collecting seaborn
  Downloading seaborn-0.13.2-py3-none-any.whl.metadata (5.4 kB)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from seaborn) (2.0.1)
Requirement already satisfied: pandas>=1.2 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from seaborn) (2.2.2)
Collecting matplotlib!=3.6.1,>=3.4 (from seaborn)
  Downloading matplotlib-3.9.1-cp312-cp312-win_amd64.whl.metadata (11 kB)
Collecting contourpy>=1.0.1 (from matplotlib!=3.6.1,>=3.4->seaborn)
  Downloading contourpy-1.2.1-cp312-cp312-win_amd64.whl.metadata (5.8 kB)
Collecting cycler>=0.10 (from matplotlib!=3.6.1,>=3.4->seaborn)
  Downloading cycler-0.12.1-py3-none-any.whl.metadata (3.8 kB)
Collecting fonttools>=4.22.0 (from matplotlib!=3.6.1,>=3.4->seaborn)
  Downloading fonttools-4.53.1-cp312-cp312-win_amd64.whl.metadata (165 kB)
     ---------------------------------------- 0.0/165.9 kB ? eta -:--:--
     -- ------------------------------------- 10.2/165.9 kB ? eta -:--:--
     ------ --------------------------------- 30.7/165.9 kB 330.3 kB/s eta 0:00:01
     ------------- -------------------------- 61.4/165.9 kB 550.5 kB/s eta 0:00:01
     ------------------- -------------------- 92.2/165.9 kB 525.1 kB/s eta 0:00:01
```

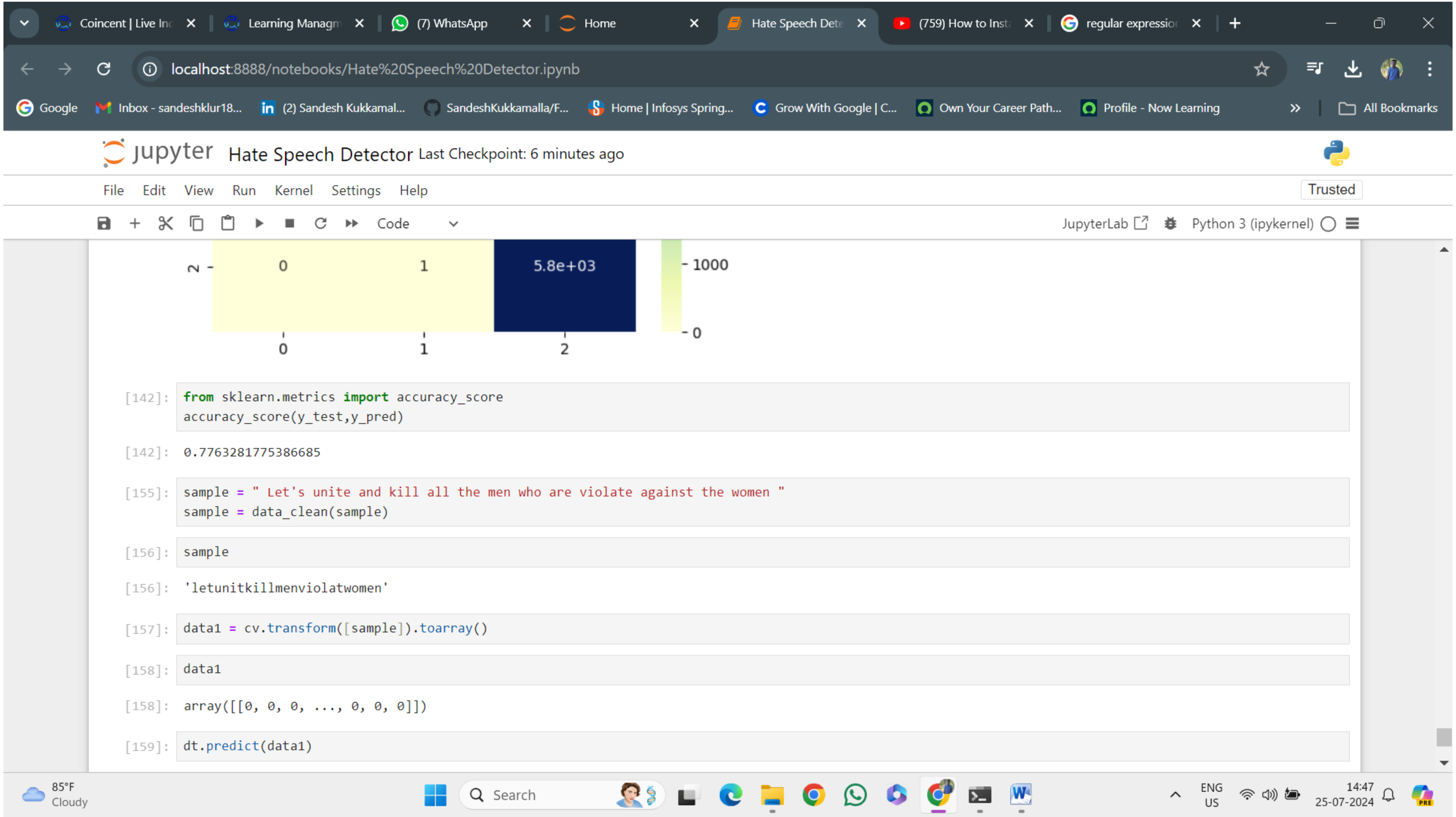


```
Requirement already satisfied: packaging>=20.0 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from matplotlib) (24.1)
Requirement already satisfied: pillow>=8 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from matplotlib) (10.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from matplotlib) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from matplotlib) (2.9.
0.post0)
Requirement already satisfied: six>=1.5 in c:\users\lenovo\appdata\local\programs\python\python312\lib\site-packages (from python-dateutil>=2.7->matplotl
ib) (1.16.0)

[notice] A new release of pip is available: 24.0 -> 24.1.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

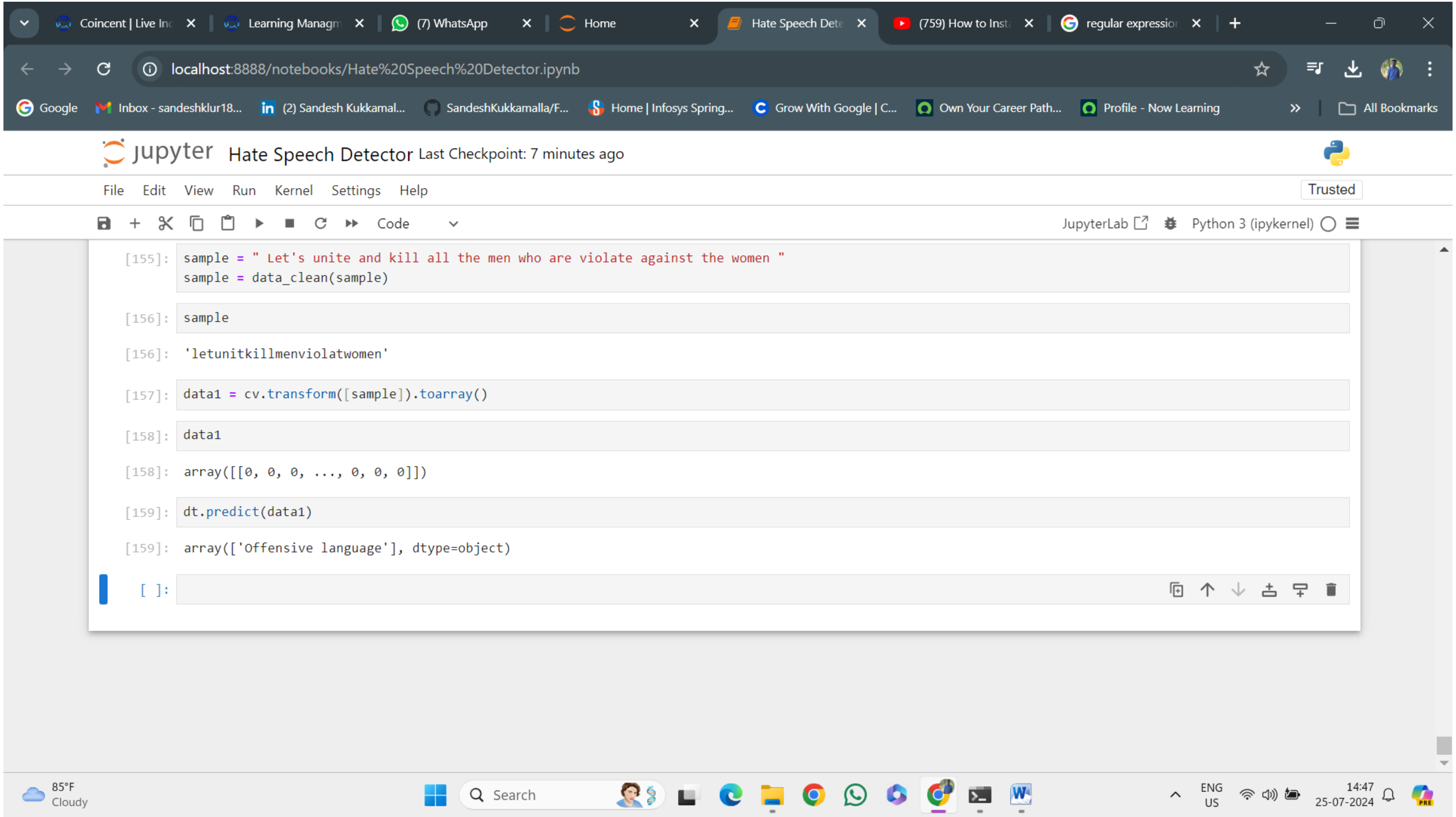
```
[131]: import matplotlib.pyplot as ply
       %matplotlib inline
```

```
[141]: sns.heatmap(cm,annot = True,cmap = "YlGnBu")
```

```
[141]: <Axes: >
```

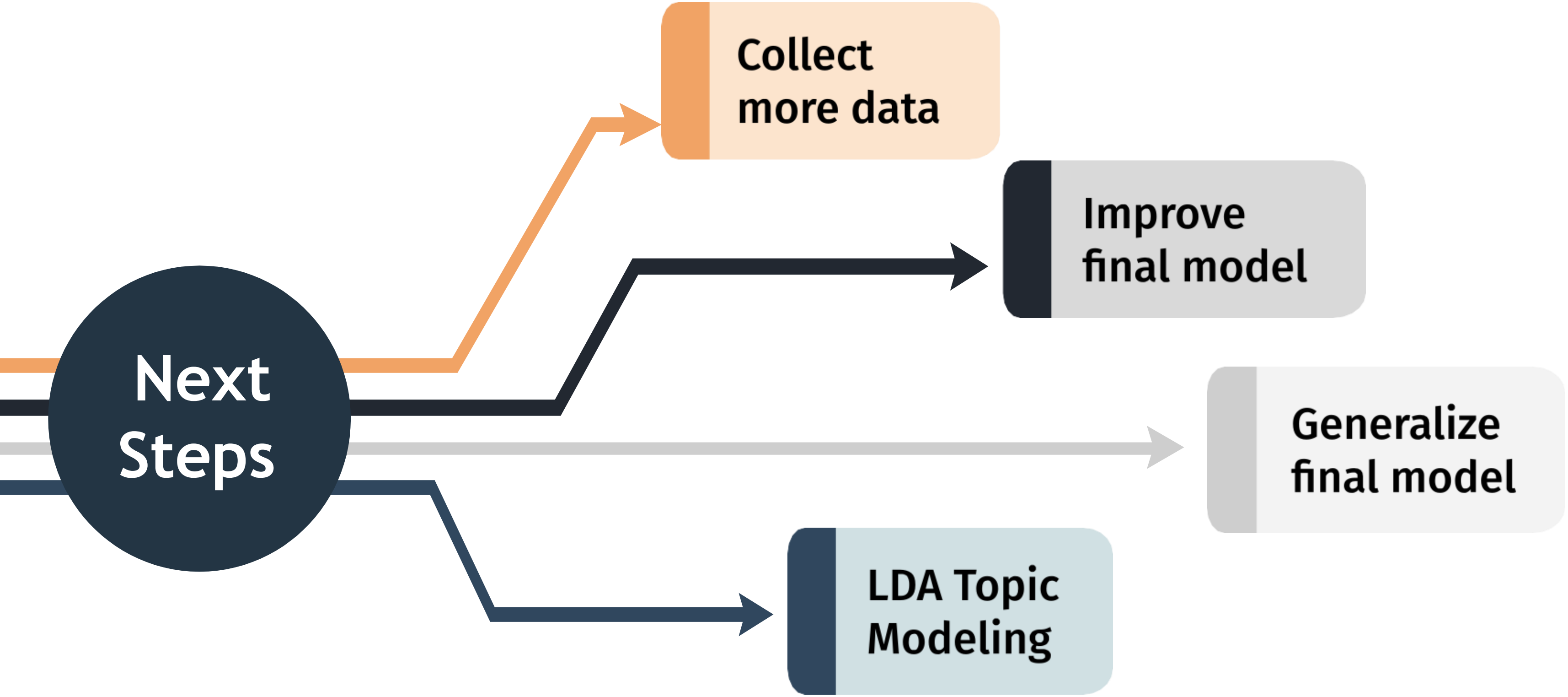


```
[142]: from sklearn.metrics import accuracy_score
       accuracy_score(y_test,y_pred)

[142]: 0.7763281775386685

[155]: sample = " Let's unite and kill all the men who are violate against the women "
       sample = data_clean(sample)

[156]: sample

[156]: 'letunitkillmenviolatwomen'

[157]: data1 = cv.transform([sample]).toarray()

[158]: data1

[158]: array([[0, 0, 0, ..., 0, 0, 0]])

[159]: dt.predict(data1)
```



```
[155]: sample = " Let's unite and kill all the men who are violate against the women "
       sample = data_clean(sample)

[156]: sample

[156]: 'letunitkillmenviolatwomen'

[157]: data1 = cv.transform([sample]).toarray()

[158]: data1

[158]: array([[0, 0, 0, ..., 0, 0, 0]])

[159]: dt.predict(data1)

[159]: array(['Offensive language'], dtype=object)

[ ]:
```

CONCLUSION


Next Steps
Collect more data
Improve final model
Generalize final model
LDA Topic Modeling

Thank You!