Filtering Profanity and Duplicates

In this phase, we will filter profanity and duplicates from the text. This is important to do because profanity can be offensive to some people, and duplicates can make the text look messy and difficult to read.We are going to use spacy and profanity filter to remove the offensive words and replace it with regular expressions. Duplicate records are time consuming and can affect proper analysis and hence need to be dropped from the dataset.

In [1]:
```
#importing basic package
import pandas as pd
```

In [2]:
```
# loading the dataset
data = pd.read_excel('Data for AI Assignment.xlsx')
data.head()
```

Out[2]:

	Text	Classification
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger

In [3]:
```
#examining the dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18001 entries, 0 to 18000
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Text            18001 non-null  object
 1   Classification  18001 non-null  object
dtypes: object(2)
memory usage: 281.4+ KB
```

In [4]:
```
# Identifying duplicate records
duplicates = data[data.duplicated(subset="Text")]
duplicates
```

Out[4]:

	Text	Classification
5067	i feel on the verge of tears from weariness i ...	joy
6133	i still feel a craving for sweet food	love
6563	i tend to stop breathing when i m feeling stre...	anger
7623	i was intensely conscious of how much cash i h...	sadness
7685	im still not sure why reilly feels the need to...	surprise
8246	i am not amazing or great at photography but i...	love
9596	ive also made it with both sugar measurements ...	joy
9687	i had to choose the sleek and smoother feel of...	joy
9769	i often find myself feeling assaulted by a mul...	sadness
9786	i feel im being generous with that statement	joy
10117	i feel pretty tortured because i work a job an...	fear
10581	i feel most passionate about	joy
11273	i was so stubborn and that it took you getting...	joy
11354	i write these words i feel sweet baby kicks fr...	love
11525	i feel a remembrance of the strange by justin ...	fear
11823	i have chose for myself that makes me feel ama...	joy
12441	i still feel completely accepted	love
12562	i feel so weird about it	surprise
12892	i cant escape the tears of sadness and just tr...	joy
13236	i feel like a tortured artist when i talk to her	anger
13846	i feel more adventurous willing to take risks ...	joy
13880	i feel like i am very passionate about youtube...	love
14107	i feel kind of strange	surprise
14314	i could feel myself hit this strange foggy wall	surprise
14634	i feel pretty weird blogging about deodorant b...	fear
14926	i resorted to yesterday the post peak day of i...	fear
15315	i will feel as though i am accepted by as well...	joy
15329	i shy away from songs that talk about how i fe...	joy
15572	i bet taylor swift basks in the knowledge that...	anger
15705	i began to feel accepted by gaia on her own terms	joy
15876	i was sitting in the corner stewing in my own ...	anger
16000	i feel like this was such a rude comment	anger
16263	i realized what i am passionate about helping ...	joy
16266	i feel so blessed and honored that we get to b...	love
16354	i could feel his breath on me and smell the sw...	joy
16416	i loved the feeling i got during an amazing sl...	joy
16503	i am feeling stressed and more than a bit anxious	anger
16587	i found myself feeling inhibited and shushing ...	sadness
16918	i feel the need to pimp this since raini my be...	joy
16960	i feel cared for and accepted	love
17027	i have not conducted a survey but it is quite ...	sadness
17276	i feel so weird and scattered with all wonders...	surprise
17888	i feel like some of you have pains and you can...	joy

In [5]:
```
#getting the length of the duplicates
duplicate_cnt=len(duplicates)
print("There are",duplicate_cnt,"duplicate records.")
```

```
There are 43 duplicate records.
```

In [6]:
```
#the list of deleted records
removed_records = duplicates.to_dict(orient="records")
removed_records
```

Out[6]:
```
[{'Text': 'i feel on the verge of tears from weariness i look at your sweet face and cant help but tenderly kis
s your cheeks',
  'Classification': 'joy'},
 {'Text': 'i still feel a craving for sweet food', 'Classification': 'love'},
 {'Text': 'i tend to stop breathing when i m feeling stressed',
  'Classification': 'anger'},
 {'Text': 'i was intensely conscious of how much cash i had left in my gas and food envelope and i still have w
hat i intended to save for next week which helps me not feel so stressed and scared',
  'Classification': 'sadness'},
 {'Text': 'im still not sure why reilly feels the need to be so weird',
  'Classification': 'surprise'},
 {'Text': 'i am not amazing or great at photography but i feel passionate about it',
  'Classification': 'love'},
 {'Text': 'ive also made it with both sugar measurements but i feel like cup is just too sweet for me',
  'Classification': 'joy'},
 {'Text': 'i had to choose the sleek and smoother feel of the sweet revenge made drawing and handling the blast
er a bit nicer',
  'Classification': 'joy'},
 {'Text': 'i often find myself feeling assaulted by a multitude of sense impressions',
  'Classification': 'sadness'},
 {'Text': 'i feel im being generous with that statement',
  'Classification': 'joy'},
 {'Text': 'i feel pretty tortured because i work a job and often the inspiration strikes while im at work',
  'Classification': 'fear'},
 {'Text': 'i feel most passionate about', 'Classification': 'joy'},
 {'Text': 'i was so stubborn and that it took you getting hurt for me to admit even to myself how i feel i have
n t been very considerate of you in that respect',
  'Classification': 'joy'},
 {'Text': 'i write these words i feel sweet baby kicks from within and my memory is refreshed i would do anythi
ng for this boy',
  'Classification': 'love'},
 {'Text': 'i feel a remembrance of the strange by justin aryiku falls into the latter category',
  'Classification': 'fear'},
 {'Text': 'i have chose for myself that makes me feel amazing',
  'Classification': 'joy'},
 {'Text': 'i still feel completely accepted', 'Classification': 'love'},
 {'Text': 'i feel so weird about it', 'Classification': 'surprise'},
 {'Text': 'i cant escape the tears of sadness and just true grief i feel at the loss of my sweet friend and sis
ter',
  'Classification': 'joy'},
 {'Text': 'i feel like a tortured artist when i talk to her',
  'Classification': 'anger'},
 {'Text': 'i feel more adventurous willing to take risks img src http cdn',
  'Classification': 'joy'},
 {'Text': 'i feel like i am very passionate about youtube and so id quite like to explain why i think youtube i
s the next best thing for entertainment',
  'Classification': 'love'},
 {'Text': 'i feel kind of strange', 'Classification': 'surprise'},
 {'Text': 'i could feel myself hit this strange foggy wall',
  'Classification': 'surprise'},
 {'Text': 'i feel pretty weird blogging about deodorant but im a bit of a deodorant snob and find it really har
d to find a good one',
  'Classification': 'fear'},
 {'Text': 'i resorted to yesterday the post peak day of illness when i was still housebound but feeling agitate
d and peckish for brew a beef http pic4',
  'Classification': 'fear'},
 {'Text': 'i will feel as though i am accepted by as well as comfortable being around both sides of my family',
  'Classification': 'joy'},
 {'Text': 'i shy away from songs that talk about how i feel toward god or that maybe even talk about my faithfu
l response toward god',
  'Classification': 'joy'},
 {'Text': 'i bet taylor swift basks in the knowledge that the boys she writes songs about probably feel torture
d',
  'Classification': 'anger'},
 {'Text': 'i began to feel accepted by gaia on her own terms',
  'Classification': 'joy'},
 {'Text': 'i was sitting in the corner stewing in my own muck feeling hated alone unworthy and violated',
  'Classification': 'anger'},
 {'Text': 'i feel like this was such a rude comment',
  'Classification': 'anger'},
 {'Text': 'i realized what i am passionate about helping women feel accepted and appreciated',
  'Classification': 'joy'},
 {'Text': 'i feel so blessed and honored that we get to be its parents',
  'Classification': 'love'},
 {'Text': 'i could feel his breath on me and smell the sweet scent of him',
  'Classification': 'joy'},
 {'Text': 'i loved the feeling i got during an amazing slalom run whether it was in training or in a race',
  'Classification': 'joy'},
 {'Text': 'i am feeling stressed and more than a bit anxious',
  'Classification': 'anger'},
 {'Text': 'i found myself feeling inhibited and shushing her quite a lot',
  'Classification': 'sadness'},
 {'Text': 'i feel the need to pimp this since raini my beloved rocky casting director loves it so much',
  'Classification': 'joy'},
 {'Text': 'i feel cared for and accepted', 'Classification': 'love'},
 {'Text': 'i have not conducted a survey but it is quite likely that many of them feel as assaulted by onel s d
emons and other creators as i would have felt had the walls been covered only with eminent figures patriotic he
roes and epic deeds',
  'Classification': 'sadness'},
 {'Text': 'i feel so weird and scattered with all wonders about a million different things',
  'Classification': 'surprise'},
 {'Text': 'i feel like some of you have pains and you cannot imagine becoming passionate about the group or the
idea that is causing pain',
  'Classification': 'joy'}]
```

In [7]:
```
# Removing duplicate records based on the "Text" column
data.drop_duplicates(subset="Text", inplace=True)

data.head()
```

Out[7]:

	Text	Classification
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger

In [8]:
```
#examining the new dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17958 entries, 0 to 18000
Data columns (total 2 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Text            17958 non-null  object
 1   Classification  17958 non-null  object
dtypes: object(2)
memory usage: 420.9+ KB
```

In [69]:
```
#importing profanity removal package

import spacy
from profanityfilter import ProfanityFilter
import re
```

In [94]:
```
# Function to filter and clean profane words from text

def filter_and_clean_profanities_and_return_detected_profanities(text):

    censored_text = profanity_filter.censor(text)


    detected_profanities = profanity_filter.is_profane(text)

    return censored_text, detected_profanities
```

In [95]:
```
# Example phrases to test the code.

example_phrases = [
  "This is a sample sentence.",
  "This sentence contains a profanity: damn."
]

# Filter the example phrases, returning the cleaned words.
filtered_phrases = [filter_and_clean_profanities_and_return_detected_profanities(phrase) for phrase in example_

# Printing the cleaned words.
for phrase in filtered_phrases:
  print(phrase)
```

```
('This is a sample sentence.', False)
('This sentence contains a profanity: ****.', True)
```

In [96]:
```
# Applying the function to the 'Text' column
data['Filtered_Text'] = data['Text'].apply(filter_and_clean_profanities_and_return_detected_profanities)
data['Filtered_Text'].head()
```

Out[96]:
```
0                    (i didnt feel humiliated, False)
1    (i can go from feeling so hopeless to so damne...
2    (im grabbing a minute to post i feel greedy wr...
3    (i am ever feeling nostalgic about the firepla...
4                        (i am feeling grouchy, False)
Name: Filtered_Text, dtype: object
```

In [121…
```
#Making a function to clean the final output

def clean_text(text):

    text = text.replace('(', '').replace(')', '')


    text = re.sub(r', (False|True)', '', text)


    text = text.replace('****', '')

    return text
```

In [123…
```
# Example usage
original_text = "(I feel like I have a ton of catching up to do, False)"
cleaned_text = clean_text(original_text)
cleaned_text
```

Out[123]:
```
'I feel like I have a ton of catching up to do'
```

In [124…
```
# Applying the function to the 'Text' column
data['Filtered_Text'] = data['Filtered_Text'].apply(clean_text)
data['Filtered_Text'].head()
```

Out[124]:
```
0                    i didnt feel humiliated
1    i can go from feeling so hopeless to so damne...
2    im grabbing a minute to post i feel greedy wr...
3    i am ever feeling nostalgic about the firepla...
4                          i am feeling grouchy
Name: Filtered_Text, dtype: object
```

In [125…
```
# Saving the filtered DataFrame to a new CSV file
data.to_csv('filtered_data.csv', index=False)
```