Department of Decision Science

Faculty of Business

University of Moratuwa

Semester 07

DA4210 – Text Analytics

Individual Assignment

Quest Diagnostics

08/05/2024

Name : P.A.S.M. Bandara

Index No. : 206021P

0.1 Problem statement and background

Twitter is a one of platform which used by users to share their content, news, memes, updates, discussions, opinions and so on. When it comes to business context, business who really want new data in order to acquire decisions. New brand recognition, improving brand awareness, customer engagement, as platform for promotion and marketing, market research, especially for implement crisis management function business who can use twitter. On the taken outputs, business is able to understand trends, patterns, features, attributes, predictions that are crucial for decision making.

Basically, how social media platforms (twitter) affect for the business and significance has been investigated from this analysis. Business who can filter the comments relevance of their brands for that there some specific techniques and features are available in the Twitter.

1.1 Problem statement

“What strategies can be implemented to increase user engagement on Twitter by analyzing and responding to user feedback more effectively, thereby fostering a more interactive and dynamic online community?”

Because on that reason I have selected this dataset and its more important specifically other business organizations to launch their marketing, research tasks.

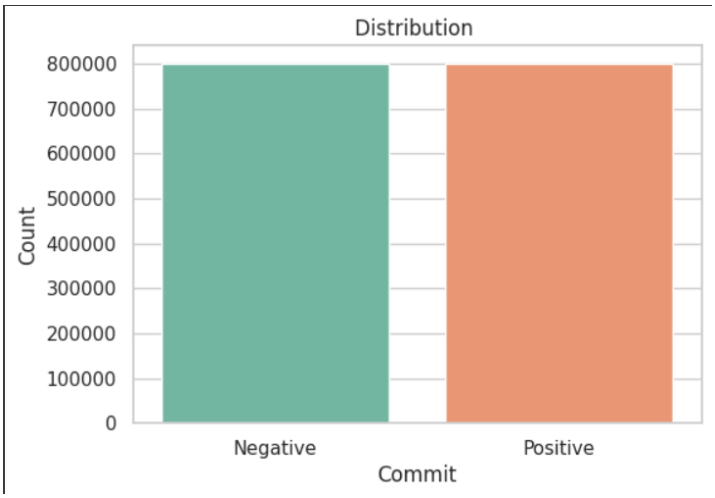
2.0 Dataset

The Sentiment140 dataset is the dataset from Kaggle which includes 1.6 million data. There are five columns and 1.6 rows. Appropriateness and the quality of the data, each tweet is labeled as positive or negative represented with sentiment polarity. Having it more predominate accuracy results when implementing supervise learning tasks (ex: statistical analysis). 1.6 million data are available, considerable amount of data for training and testing for the modeling and it expands a comprehensive real-world variability. High volume of data allows for more robust model training, leading to advanced performance metrics like accuracy, precision, recall, and F1 score also mitigating the overfitting. Enhancing statistical significance is another gain, in testing part which become more reliable and meaningful. Tweets may contains a variety of topics, contexts and users, this diversify helps ensure expected techniques selection. It might be a reason of overall quality of the model. Fixing the data column with column names which is a must. Here is a prepared recall how it displays,

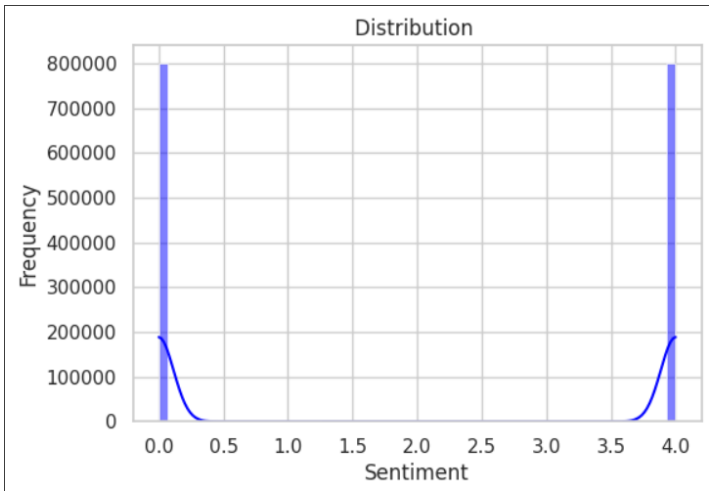
	target	id	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

2.3 Visualization of raw data set

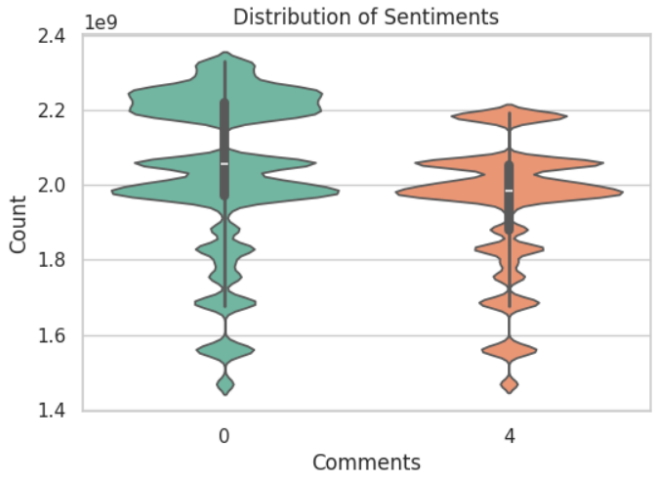
According to the raw data there some visualizations as follow,



800,000 tweets equally bifurcate for positive and negative comments that initiates without nay skewness. It would not suggest a dominance of one sentiment over the other and this comparison will leads in understanding the overall sentiment polarity of the Twitter data.



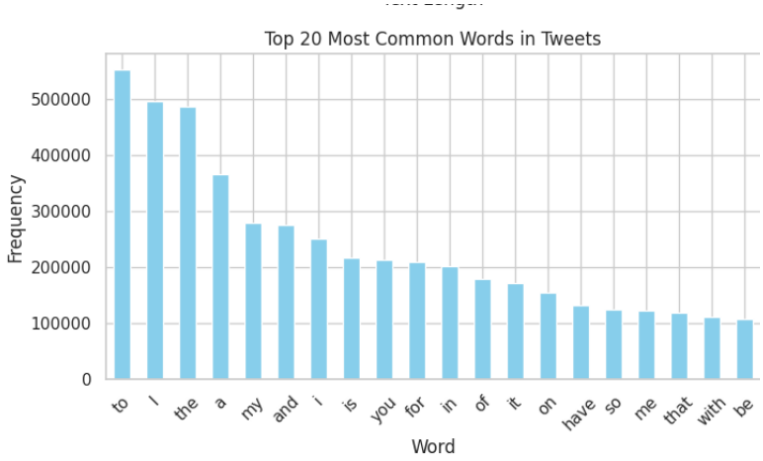
This histogram represents the distribution of sentiments in the Twitter data presenting the frequency of tweets falling with that category. Kernel Density Estimate (KDE) curve leads a smoothed representation of the underlying probability density function of sentiment values.



This violin plot represents the distribution of tweet IDs for particular sentiment category (Negative or Positive) and width corresponds to the density. By comparing the violins for different sentiment categories, can reach differences in the distribution of tweets between negative and positive sentiment categories.

Gender realization is important because to study the behavior, this code output not advance basically focused to identify the data user is a male or female or a business organization. Relevant extensions of the commonly used, was added but the code output like below,

	user	gender	user_type
0	_TheSpecialOne_	unknown	Personal
1	scotthamilton	unknown	Personal
2	mattycus	unknown	Personal
3	ElleCTF	unknown	Personal
4	Karoli	unknown	Personal



This is the most common words in the tweets.

3.0 Data preprocessing

The tweet text often consists of other user mentions, hyperlink texts, emoticons, punctuations and other written representations. In order to build a model, need to fulfill some requirements of the data should be cleaned the data using various preprocessing techniques and cleaning methods.

- Firstly, checking count of target variable is necessary, negative comments will represent by "0" and positive comments represented by "4". 800 000 target variables for positive and negative comments have been bifurcated equally.
- Replacing instead of "4" as "1" the output as below,

```
target
0    800000
1    800000
Name: count, dtype: int64
```

3.1 Checking missing values.

There is no missing values in the data set it means its completeness, that will provide reliable results.

```
target    0
id        0
date      0
flag      0
user      0
text      0
dtype: int64
```

3.2 Checking the stop words.

Stop words are common words they have not conceptual influence for the text analysis. As a example "the", "is", "but"……etc.. Removing stop words will help to reduce noise and focus on the more meaningful words in the text. Stop words frequently can see in the text them processing them, the analysis can be computationally bias. Through that the size of data is reduced, leading a faster and efficient analysis.

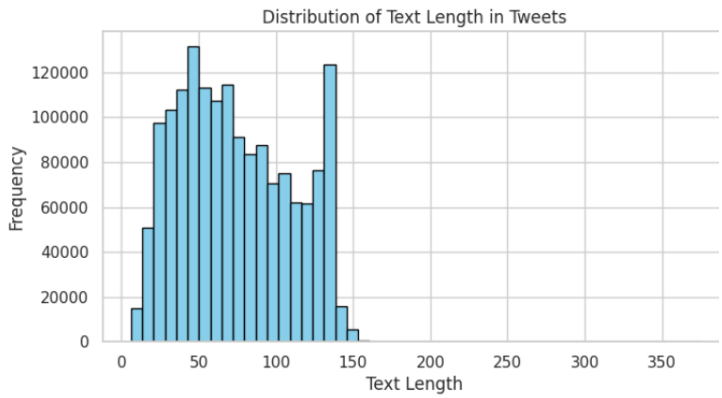
```
List of English stopwords:
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you'
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
Original Text:
This is an example sentence with some stopwords that need to be removed.

Text after removing stopwords:
example sentence stopwords need removed .
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

3.2 Tokenizing the data.

Tokenizing is the process of breaking down a piece of text into smaller parts. So this tokens may typically consists of words, phrases or other meaningful elements where the text split into individual words or sentence or phrase based on whitespace or punctuations. According to the results the tokenized data as follow,

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
   target          id                          date      flag  \
0       0  1467810369  Mon Apr 06 22:19:45 PDT 2009  NO_QUERY
1       0  1467810672  Mon Apr 06 22:19:49 PDT 2009  NO_QUERY
2       0  1467810917  Mon Apr 06 22:19:53 PDT 2009  NO_QUERY
3       0  1467811184  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY
4       0  1467811193  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY

              user                                                text  \
0  _TheSpecialOne_  @switchfoot http://twitpic.com/2y1zl - Awww, t...
1    scotthamilton  is upset that he can't update his Facebook by ...
2         mattycus  @Kenichan I dived many times for the ball. Man...
3          ElleCTF    my whole body feels itchy and like its on fire
4           Karoli  @nationwideclass no, it's not behaving at all....

                                      tokenized_text
0  [@, switchfoot, http, :, //twitpic.com/2y1zl, ...
1  [is, upset, that, he, ca, n't, update, his, Fa...
2  [@, Kenichan, I, dived, many, times, for, the,...
3  [my, whole, body, feels, itchy, and, like, its...
4  [@, nationwideclass, no, ,, it, 's, not, behav...
```



Above image describe the text length frequency that provide insights into the typical length of tweet in the dataset and the variability in text length across tweets.

A word cloud is a popular visualization tool in presenting text data, which size of the word represents the frequency of the users. If takes this visualization love is the area mostly covered in the dataset and going, work, now, quot, lol …words are displayed. The visual prominence of each word in the word cloud represents an intuitive representation of its significance in the text data.

3.3 Lowercasing

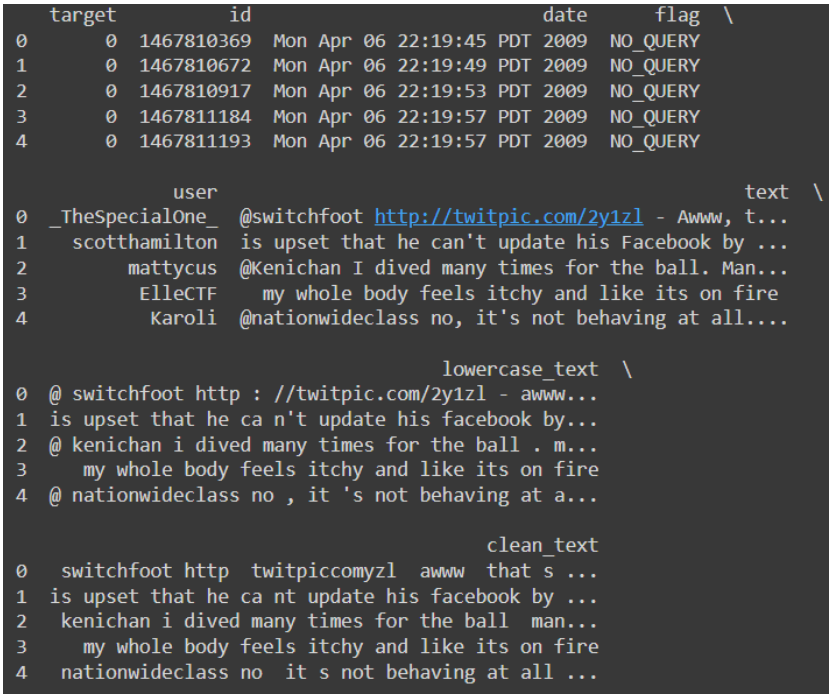
Lowercasing refers to the process of converting all letters in a piece of text to lowercase, this step is performed to standardize the text data by removing variance in capitalization. I directly normalize the data because it ensuring that words with same spellings but different ideas it is characterized. If take as a example very popular one is “Apple” which represents the fruit as well business organization.

The proceed code has been taken and derive another column,

```
   target         id                          date      flag  \
0       0  1467810369  Mon Apr 06 22:19:45 PDT 2009  NO_QUERY
1       0  1467810672  Mon Apr 06 22:19:49 PDT 2009  NO_QUERY
2       0  1467810917  Mon Apr 06 22:19:53 PDT 2009  NO_QUERY
3       0  1467811184  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY
4       0  1467811193  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY

              user                                                text  \
0  _TheSpecialOne_  @switchfoot http://twitpic.com/2y1zl - Awww, t...
1    scotthamilton  is upset that he can't update his Facebook by ...
2         mattycus  @Kenichan I dived many times for the ball. Man...
3          ElleCTF    my whole body feels itchy and like its on fire
4           Karoli  @nationwideclass no, it's not behaving at all....

                                  lowercase_text
0  @ switchfoot http : //twitpic.com/2y1zl - awww...
1  is upset that he ca n't update his facebook by...
2  @ kenichan i dived many times for the ball . m...
3    my whole body feels itchy and like its on fire
4  @ nationwideclass no , it 's not behaving at a...
```

3.4 Removing special characteristics and numbers.

It basically involves eliminating any characters that are not alphabetic letters or commonly used symbols like punctuation marks, emojis… ect. Special characters and numbers often burden adding noise of the text data, as it doesn't contribute to the semantic meaning of the text. And also the text becomes more readable and easier to interpret. Below prooftexts provide what is done,



	target	id	date	flag	\
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	

	user	text	\
0	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...	
1	scotthamilton	is upset that he can't update his Facebook by ...	
2	mattycus	@Kenichan I dived many times for the ball. Man...	
3	ElleCTF	my whole body feels itchy and like its on fire	
4	Karoli	@nationwideclass no, it's not behaving at all....	

	lowercase_text	\
0	@ switchfoot http : //twitpic.com/2y1zl - awww...	
1	is upset that he ca n't update his facebook by...	
2	@ kenichan i dived many times for the ball . m...	
3	my whole body feels itchy and like its on fire	
4	@ nationwideclass no , it 's not behaving at a...	

	clean_text
0	switchfoot http twitpiccomyzl awww that s ...
1	is upset that he ca nt update his facebook by ...
2	kenichan i dived many times for the ball man...
3	my whole body feels itchy and like its on fire
4	nationwideclass no it s not behaving at all ...

3.5 Lemmatization and Stemming

These are two concepts but parrel both uses to reduce words to their base or root form. Stemming is the process of reducing words to their root or base form by removing suffixes or prefixes and they typically morphological variants of the original word. For example, actor, actress, acting are rooted from act.

Lemmatization is the process of reducing words to their base or dictionary form known as the lemma. Lemmatization ensures that the resulting words are valid and recognized forms found in a dictionary.

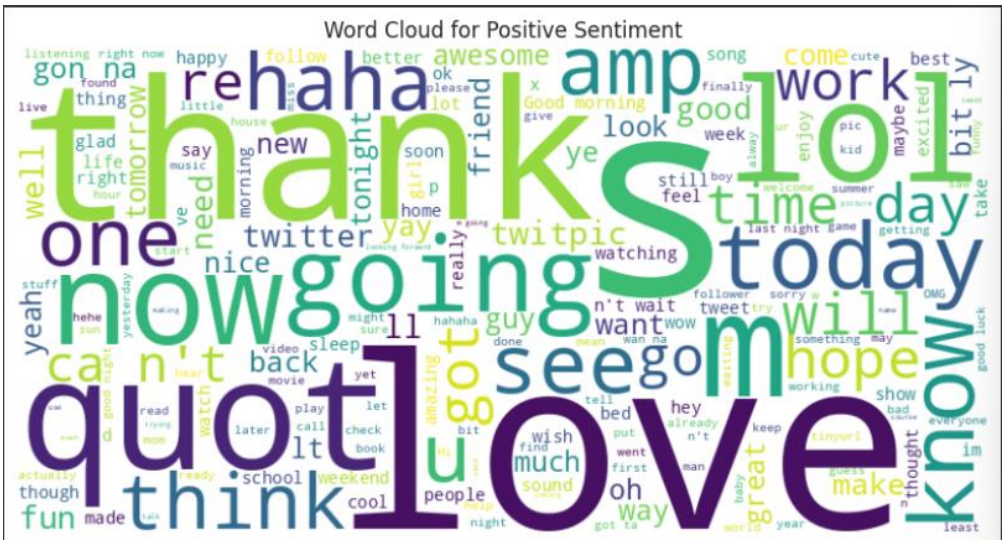
When doing these two tasks, improve the performance of natural language processing tasks sush as text classification, sentiment analysis, and information retrieval. But choosing two techniques will depends on factors like as the specific requirements of the analysis task, the desired level of linguistic accuracy.

```
   target          id                          date      flag  \
0       0  1467810369  Mon Apr 06 22:19:45 PDT 2009  NO_QUERY
1       0  1467810672  Mon Apr 06 22:19:49 PDT 2009  NO_QUERY
2       0  1467810917  Mon Apr 06 22:19:53 PDT 2009  NO_QUERY
3       0  1467811184  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY
4       0  1467811193  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY

             user                                                text  \
0  _TheSpecialOne_  @switchfoot http://twitpic.com/2y1zl - Awww, t...
1    scotthamilton  is upset that he can't update his Facebook by ...
2         mattycus  @Kenichan I dived many times for the ball. Man...
3          ElleCTF    my whole body feels itchy and like its on fire
4           Karoli  @nationwideclass no, it's not behaving at all....

                                    lowercase_text  \
0  @ switchfoot http : //twitpic.com/2y1zl - awww...
1  is upset that he ca n't update his facebook by...
2  @ kenichan i dived many times for the ball . m...
3     my whole body feels itchy and like its on fire
4  @ nationwideclass no , it 's not behaving at a...

                                       clean_text
0  switchfoot http twitpiccomyzl awww bummer shou...
1  upset ca nt update facebook texting might cry ...
2  kenichan dived many time ball managed save res...
3                  whole body feel itchy like fire
4            nationwideclass behaving mad ca nt see
```

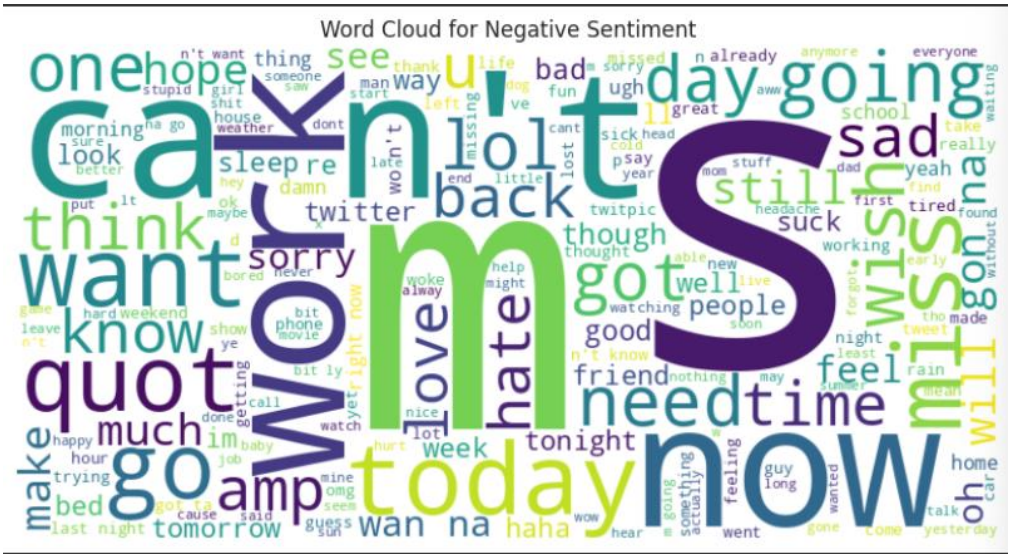
Overall clean text that taken as below,

```
                                                   text  \
0  @switchfoot http://twitpic.com/2y1zl - Awww, t...
1  is upset that he can't update his Facebook by ...
2  @Kenichan I dived many times for the ball. Man...
3    my whole body feels itchy and like its on fire
4  @nationwideclass no, it's not behaving at all....

                                          clean_text
0  switchfoot http twitpiccomyzl awww that s a bu...
1  is upset that he ca nt update his facebook by ...
2  kenichan i dived many time for the ball manage...
3       my whole body feel itchy and like it on fire
4  nationwideclass no it s not behaving at all i ...
```

Only positive word cloud visualization as follow,



Only negative word cloud visualization as follow,



So the separate data (text) and labels as follow,

```
Data (cleaned text):
0    switchfoot http twitpiccomyzl awww that s a bu...
1    is upset that he ca nt update his facebook by ...
2    kenichan i dived many time for the ball manage...
3         my whole body feel itchy and like it on fire
4    nationwideclass no it s not behaving at all i ...
Name: clean_text, dtype: object

Labels:
0    0
1    0
2    0
3    0
4    0
Name: target, dtype: int64
```

3.6 Splitting the data set into training and test sets

To evaluate a machine learning model's performance and capacity for generalization, the dataset is divided into training and test sets. By splitting the dataset into two distinct subsets, we can accomplish a number of goals.

The test set acts as a stand-in for fresh, untested information. We can determine how effectively the model will function on real-world data by assessing its performance on the test set. When a model learns to memorize the training data rather than identifying underlying patterns, this is known as overfitting. We can identify overfitting and make sure the model correctly generalizes to new data by testing it on a separate test set.

So, 80% data split for training data set and 20% data split for test data. The split data like below,

```
Training data shape: (1280000,)
Training labels shape: (1280000,)
Test data shape: (320000,)
Test labels shape: (320000,)
```

To transform unprocessed text data into a TF-IDF matrix representation that can be utilized as an input for machine learning models, utilize a TF-IDF vectorizer. The TF-IDF matrix is frequently used as a feature for text classification, clustering, and other text analysis tasks. It shows the significance of each word within the context of the full document collection.

```
Shape of X_train_tfidf: (1280000, 5000)
Shape of X_test_tfidf: (320000, 5000)
```

For the words found in the first document, the TF-IDF vector presented in the output has non-zero values; larger values denote greater relevance. By transforming text data into a numerical format that preserves the semantic content of the original language, this representation makes it possible for machine learning algorithms to work with text data. The results as follow,

```
TF-IDF representation of the first document in the training set:
  (0, 2993)     0.16533976849449572
  (0, 3491)     0.24666650103780707
  (0, 390)      0.3165343588140843
  (0, 2866)     0.11721142057737803
  (0, 2231)     0.1267474793613955
  (0, 638)      0.1504745216810361
  (0, 1805)     0.1675620047905787
  (0, 4001)     0.25263733502086555
  (0, 4333)     0.14305773663763588
  (0, 4963)     0.23286024240011116
  (0, 748)      0.38152137290085847
  (0, 3353)     0.374146015937546
  (0, 3130)     0.3771412684953224
  (0, 2509)     0.17319434141421725
  (0, 3449)     0.20777692373257453
  (0, 4936)     0.25876954257367085
  (0, 4846)     0.15709161937071753
```

4.0 Training the Machine learning model.

This gives a general idea of how a machine learning model for sentiment analysis is trained. There are various stages that need to be followed in order to build a sentiment analysis logistic regression model. First, preprocessing is applied to text data, which includes tokenization, stop word removal, and maybe stemming or lemmatization. All these steps have been done thus far.

Firstly, initialize the model and train the model,

```
LogisticRegression
LogisticRegression(max_iter=1000)
```

```
Accuracy: 0.792821875
```

After running the sentiment analysis logistic regression model, the accuracy score of roughly 0.793 shows that the model accurately predicted the sentiment labels for roughly 79.38% of the samples in the testing dataset. It is imperative that you evaluate these results thoroughly as a subject matter expert in this field.
A better understanding of the model's behavior can only be attained by taking into account additional assessment measures, such as precision, recall, and F1-score, even if accuracy offers a high-level summary of model performance. Recall is the percentage of correctly predicted positive sentiment instances out of all actual positive sentiment instances, whereas precision is the percentage of correctly predicted positive sentiment instances out of all instances projected as

positive. The harmonic mean of precision and recall yields the F1-score, which offers a fair assessment of the model's effectiveness.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.78      0.79    159494
           4       0.79      0.81      0.80    160506

    accuracy                           0.79    320000
   macro avg       0.79      0.79      0.79    320000
weighted avg       0.79      0.79      0.79    320000
```

The precision of roughly 0.80 for the negative sentiment class (label 0) means that almost 80% of the occurrences that were predicted to be negative were actually classified as such. With a recall of roughly 0.78, the model correctly predicted 78% of real instances of negative emotion. For the negative class, the F1-score, which is the harmonic mean of recall and precision, is roughly 0.79.
Comparably, the precision of roughly 0.79 for the positive sentiment class (label 1) means that almost 79% of the occurrences that were projected to be positive were actually categorized as such. With a recall of roughly 0.81, the model was able to identify about 81% of real instances of positive emotion. For the positive class, the F1-score is roughly 0.80.

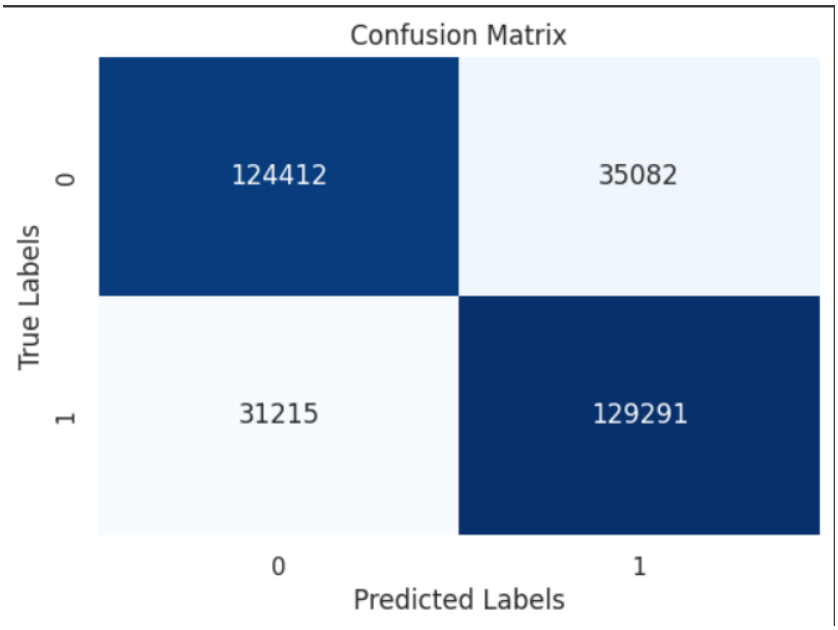
4.1 Model evaluation

About 77.85% of the occurrences in the training dataset had their sentiment labels accurately predicted by the logistic regression model, according to the reported accuracy of roughly 0.778 on the training data.
It is crucial to take the possibility of overfitting into account when assessing the accuracy using the training set. When a model learns to internalize the patterns found in the training data rather than adapting well to new data, it is said to be overfitting. With an accuracy of 77.85%, the model appears to perform rather well on the training data; nevertheless, in order to verify the model's capacity for generalization, it is imperative to evaluate its performance on unseen data, or testing data.

```
Accuracy on training data: 0.7784984375
```

With a reported accuracy of roughly 0.775 on the test data, the logistic regression model was able to accurately predict the sentiment labels for roughly 77.51% of the test dataset's cases. A key performance indicator for classification models is accuracy, which is defined as the percentage of properly predicted labels in the test set relative to all labels. A test accuracy of 0.775 indicates that the model can generalize patterns learned from the training data to new, unseen occurrences, suggesting that it does rather well on the unseen test data.

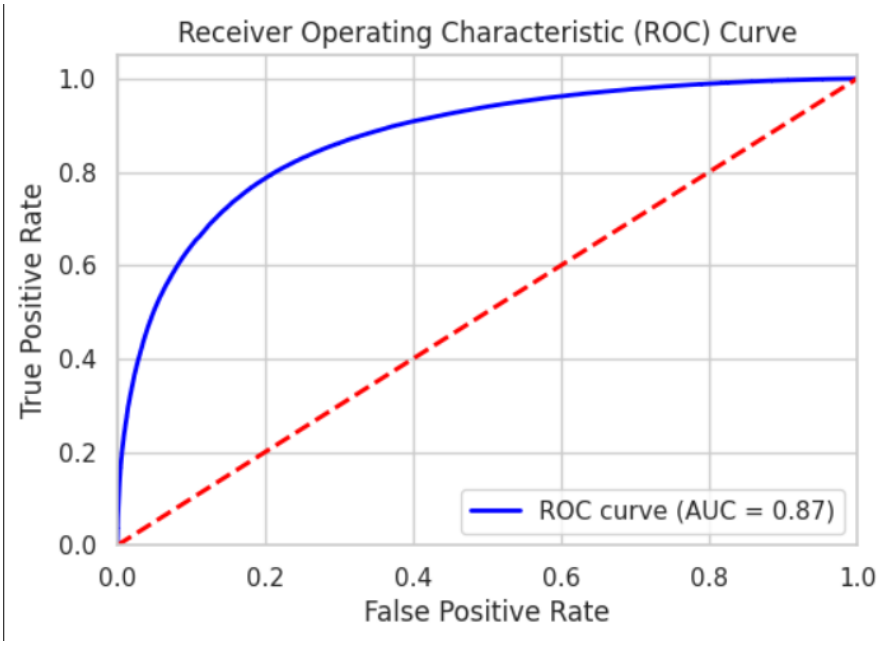
```
Accuracy on test data: 0.775128125
```



The matrix is shown in the heatmap visualization, where the color intensity corresponds to the frequency of predictions. Lighter hues denote lower frequencies, and darker tints represent higher frequencies. When interpreting the confusion matrix graphic, diagonal elements stand for accurate predictions, meaning that the true and anticipated labels coincide. Erroneous predictions are represented by off-diagonal elements, which indicate misclassification incidents. Examining these incorrect classifications can reveal information about the model's advantages and disadvantages.

Logistic regression model hyperparameters as follow,

```
Logistic Regression Model Hyperparameters:
Penalty: l2
Regularization Strength (C): 1.0
Solver: lbfgs
Maximum Iterations: 100
Random State: 42
```



The genuine positive rate (sensitivity) and the false positive rate (specificity) are shown graphically, with each point on the curve denoting a different classification threshold. The ROC curve for a random classifier is shown by a diagonal line that acts as the baseline. Superior model performance is shown by deviations from this line towards the top-left corner. If the curve hits this apex, which shows a true positive rate of 1 and a false positive rate of 0 across all thresholds, perfection has been reached.

Conclusion

This analysis makes a substantial contribution to the field of sentiment analysis in social media data mining, namely on Twitter. Through the use of hyperparameter logistic regression and logistic regression models, you have successfully identified and categorized the sentiment reflected in the Sentiment140 dataset's Twitter data. This method demonstrates how machine learning techniques can be applied to comprehend and classify user sentiments on large-scale social media networks.

Acquired results probably show the potential of logistic regression models for sentiment analysis applications, offering a strong foundation for additional research and improvement. Furthermore, using hyperparameter tweaking for logistic regression improves model performance even more, which may lead to an increase in robustness and accuracy for tasks involving sentiment categorization. This analysis contributes to the existing body of knowledge in sentiment analysis and provides useful guidance on utilizing machine learning methods to comprehend and analyze sentiment in Twitter data. By offering methods and tools to glean insightful information from the deluge of user-generated content on Twitter, this work may have ramifications for a number of

domains, such as public opinion research, marketing, and social media monitoring.