Sentiment Analysis

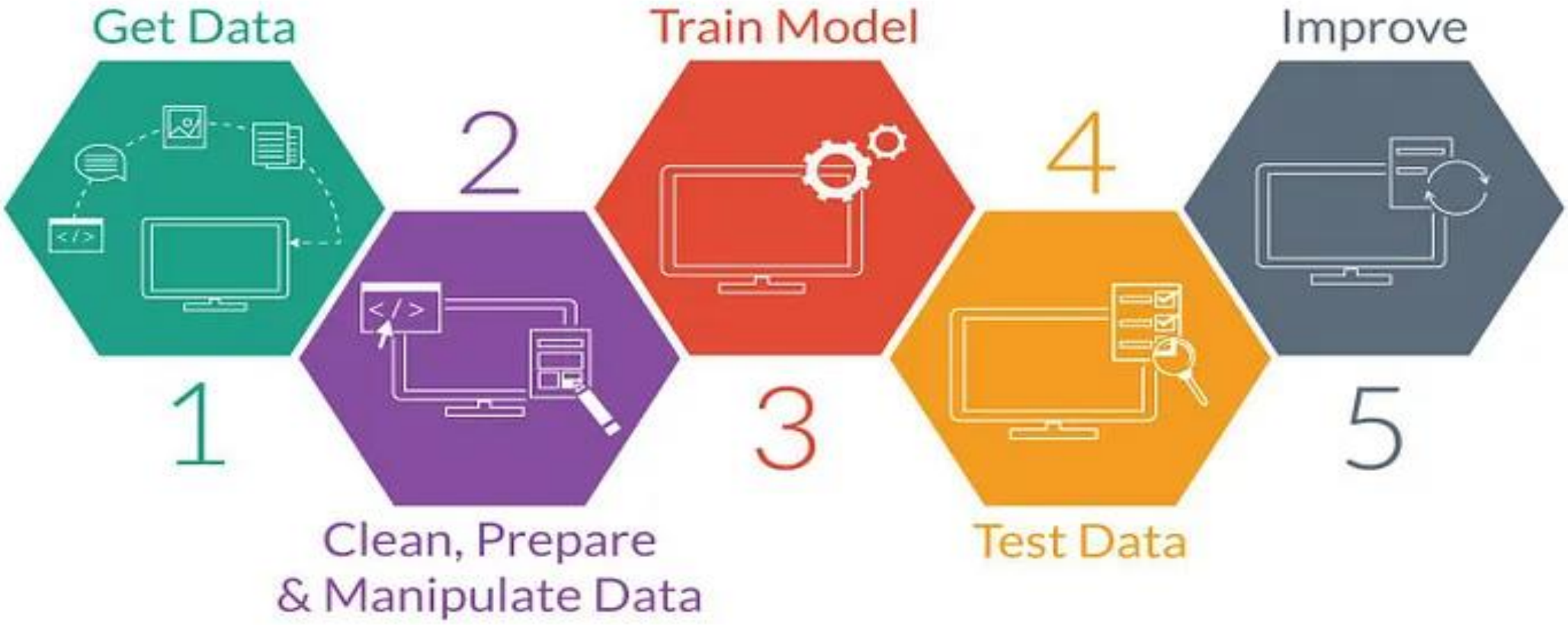
SHRADDHA SAHA

Introduction

- Sentiment analysis, a branch of natural language processing (NLP), is a fascinating field that delves into the understanding and interpretation of human emotions expressed in textual data. With the explosion of social media and online platforms, the amount of text data available for analysis has grown exponentially. Sentiment analysis has emerged as a powerful tool for extracting valuable insights from this vast reservoir of textual information.

- One of the pivotal datasets in sentiment analysis is the Sentiment140 dataset, which has played a significant role in advancing research and applications in this domain. Compiled by Alec Go, Richa Bhayani, and Lei Huang, the Sentiment140 dataset consists of over 1.6 million tweets, each labeled with sentiment polarity – positive, negative, or neutral. This dataset serves as a benchmark for sentiment analysis tasks, facilitating the development and evaluation of algorithms and models.

Workflow of the project:



For work, I have used the sentiment140 dataset taken from www.Kaggle.com.

This dataset contains 1600000 rows and 6 columns. We have inserted the column names for ease of work. The first few rows are shown below:

	target	ids	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....

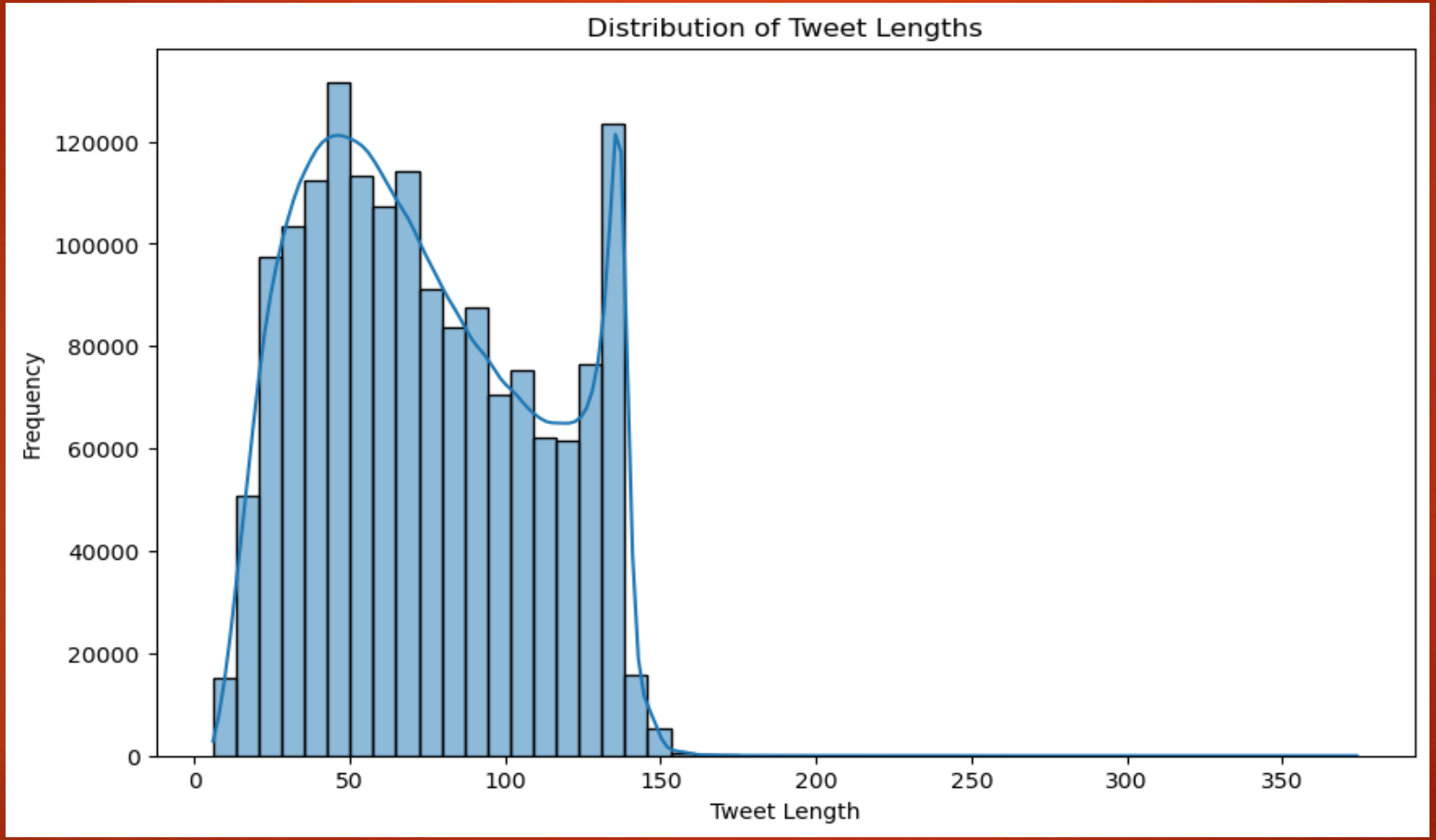
Next, I did an Exploratory Data Analysis. For, this dropped some unnecessary columns.
Then, added an extra column “sentiment” converting the target values to human-readable labels.
Then, removed neutral tweets for EDA.
And found :

- Distribution of Sentiment Classes:

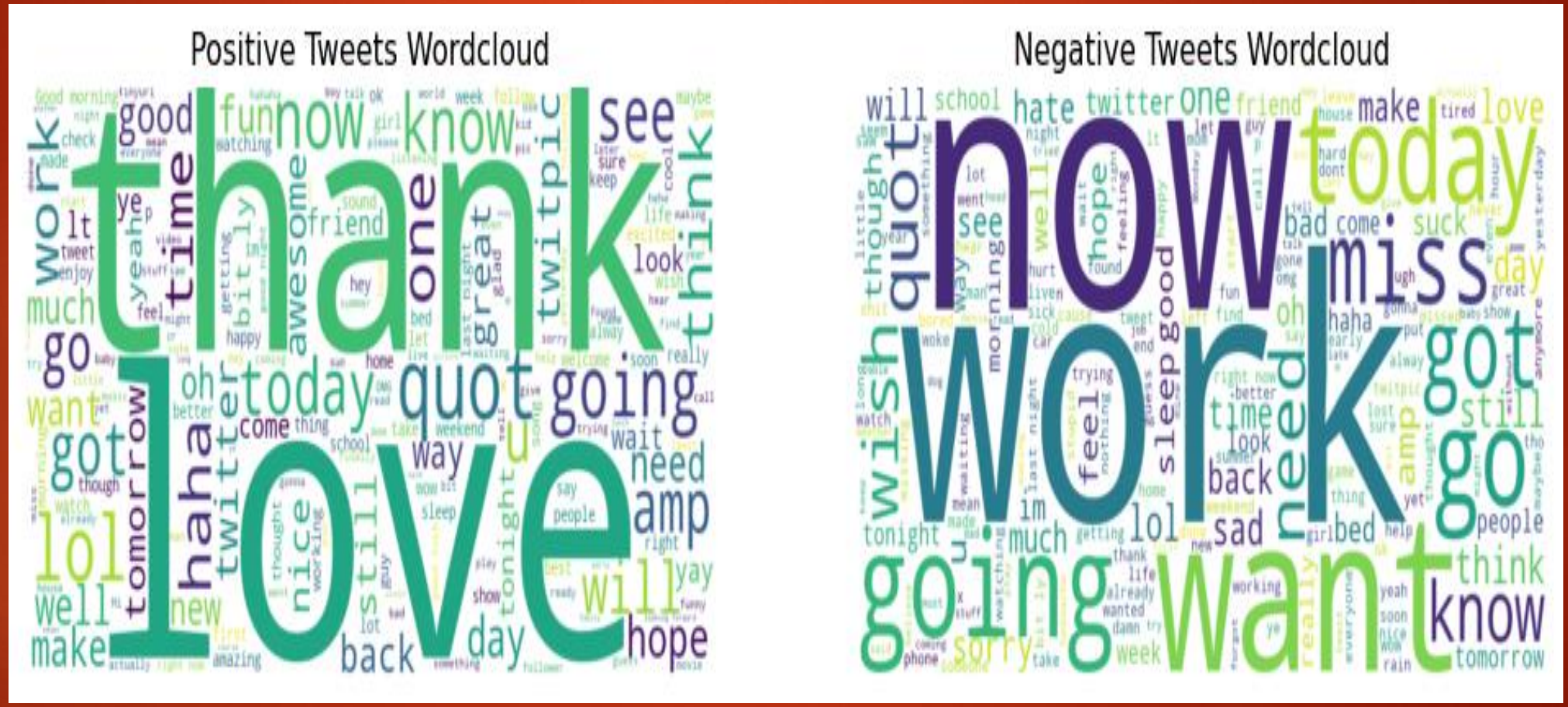
```
Negative 800000
Positive 800000

Name: sentiment, dtype: int64
```

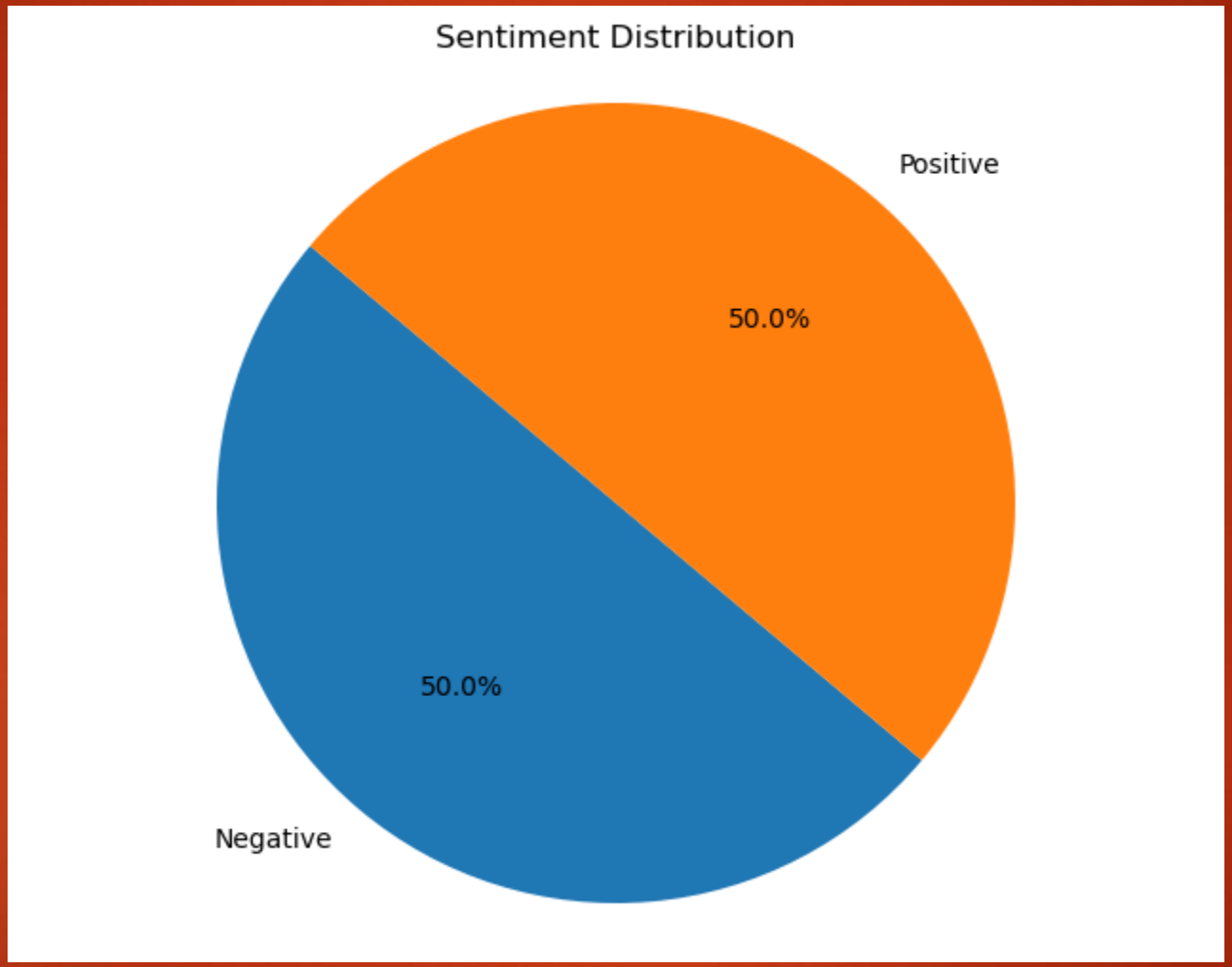
- Plotting the distribution of tweet lengths:



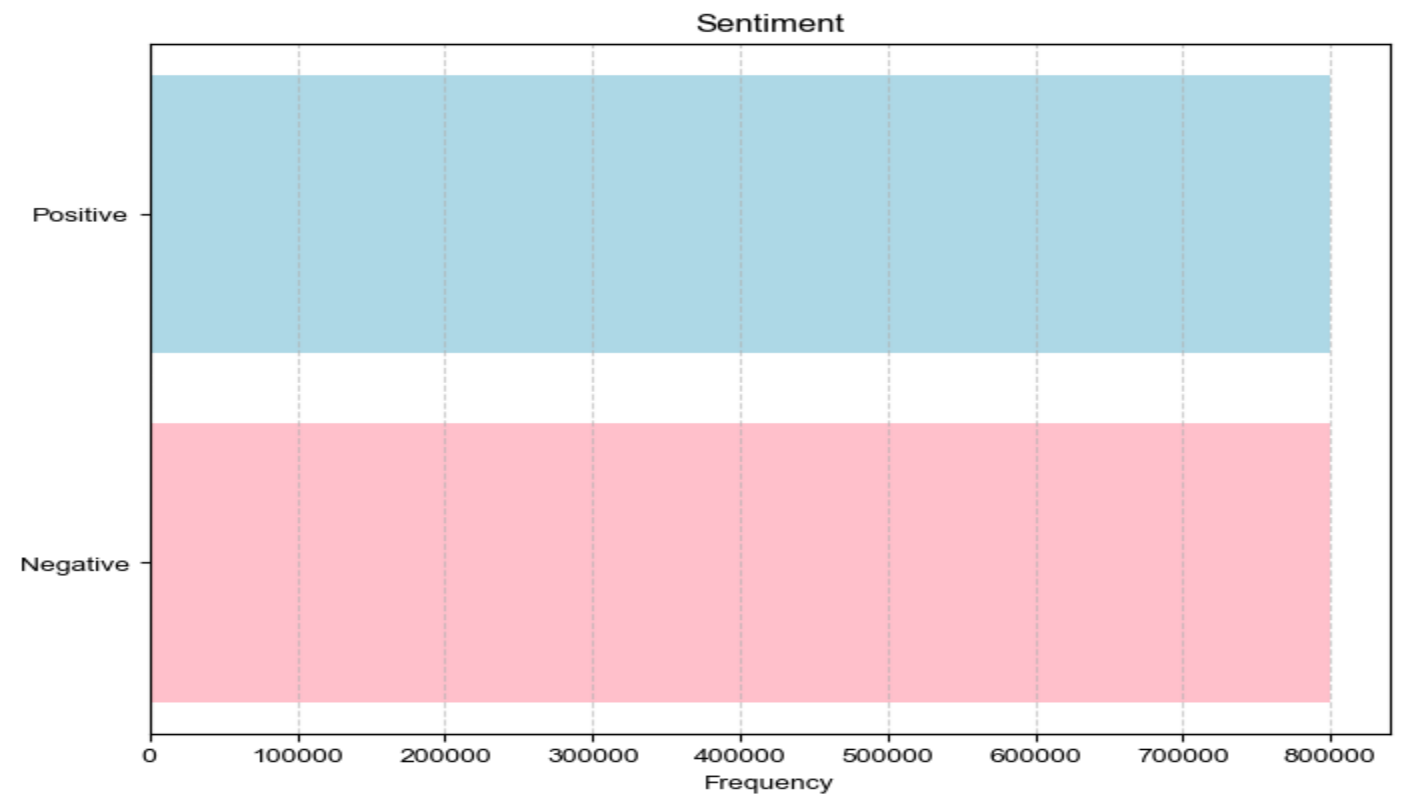
Plotting word clouds:



Pie chart for sentiment distribution:



Sentiment distribution in a horizontal barplot :



Next, done pre-processing by removing URLS, html tags, Punctuation, words that have numbers , digits, white spaces.

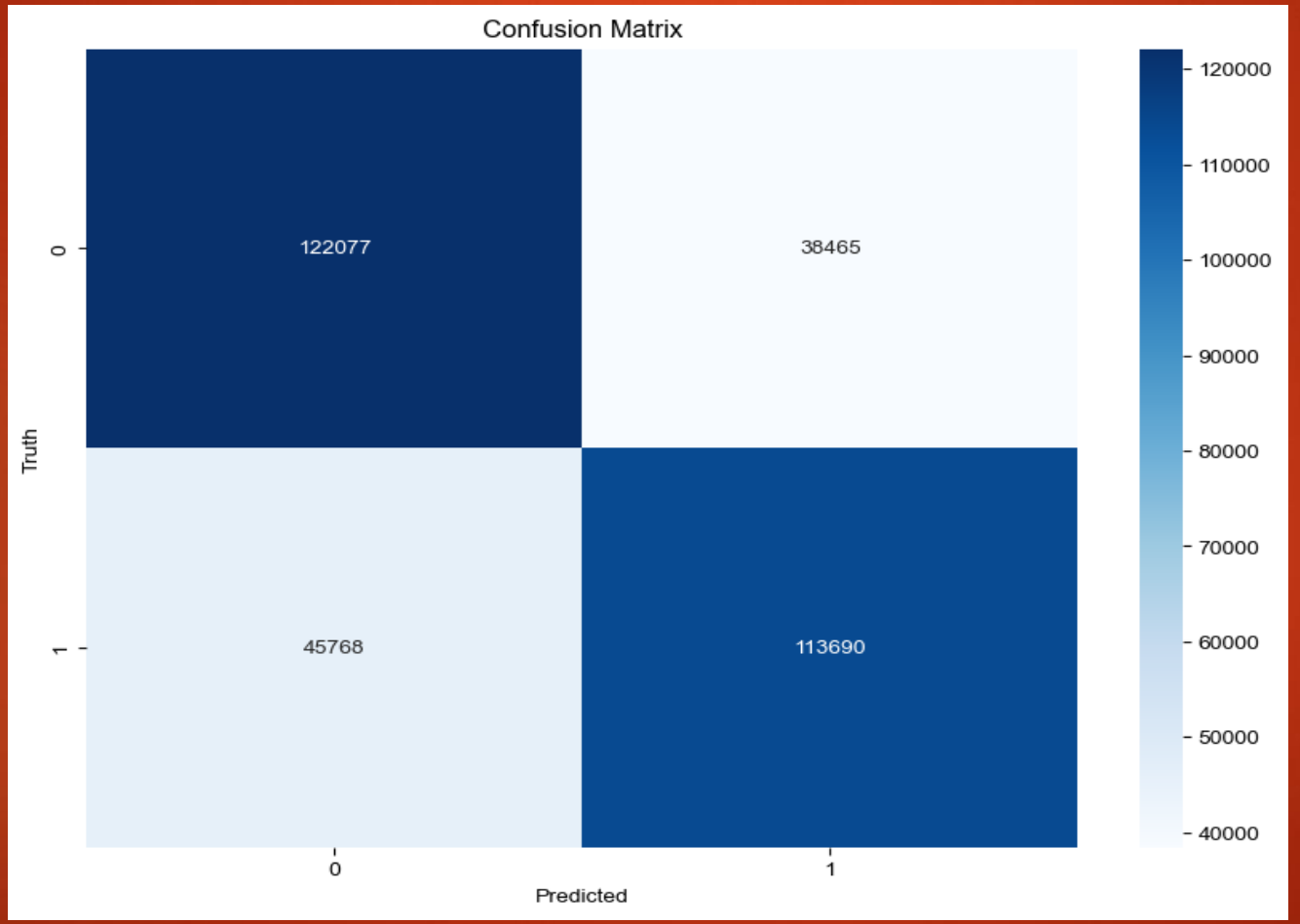
Finally, train-test split is done.

```
Train Data size: 1280000 1280000
Test Data size 320000
```

Next, done padding:

```
After padding: (1280000, 27)
After padding:(320000, 27)
```

- with batch_size=512, vocab_size=len(tokenizer.word_index) +1 embedding_dim = 100 and epoch =20, we fitted our CNN . The confusion matrix is:



Finally, the classification report is:

	precision	recall	f1-score	support
Negative	0.73	0.76	0.74	160542
Positive	0.75	0.71	0.73	159458
accuracy			0.74	320000
macro avg	0.74	0.74	0.74	320000
weighted avg	0.74	0.74	0.74	320000