Adipurush Sentiment Analysis: Harnessing Machine Learning to Understand Twitter Buzz

In [144]:

```
import pandas as pd
```

In [2]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [4]:

```
df = pd.read_csv('adipurush_tweets.csv')
```

In [5]:

```
df
```

Out[5]:

	Date Created	Number of Likes	Source of Tweet	Tweets
0	2023-06-30 09:21:00+00:00	0	NaN	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...
1	2023-06-30 09:20:57+00:00	0	NaN	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟️ \n@go...
2	2023-06-30 09:20:22+00:00	0	NaN	@ponilemova #Adipurush
3	2023-06-30 09:20:00+00:00	3	NaN	Adipurush VS 72 Hoorain VS The Kerala Story Co...
4	2023-06-30 09:15:22+00:00	3	NaN	ST: #Adipurush https://t.co/lsGKcgQuKL
...	...	...	...	...
9996	2023-06-23 10:08:50+00:00	0	NaN	Adipurush 1st Week WW Box Office Collections: ...
9997	2023-06-23 10:08:49+00:00	0	NaN	#GodMorningFriday\nवास्तव में #Adipurush यानि ...
9998	2023-06-23 10:08:17+00:00	3101	NaN	Let the empowering lyrics of #Shivoham elevate...
9999	2023-06-23 10:08:01+00:00	0	NaN	When it comes to choosing a service or product...
10000	2023-06-23 10:07:45+00:00	0	NaN	A film about #Ramayana, our greatest epic coul...

10001 rows × 4 columns

In [6]:

```
df.shape
```

Out[6]:

```
(10001, 4)
```

In [7]:

```
df.columns
```

Out[7]:

```
Index(['Date Created', 'Number of Likes', 'Source of Tweet', 'Tweets'], dtype='object')
```

In [8]:

```
df.duplicated().sum()
```

Out[8]:

```
1
```

In [9]:

```
df = df.drop_duplicates()
```

In [10]:

```
df.isnull().sum()
```

Out[10]:

```
Date Created           0
Number of Likes        0
Source of Tweet    10000
Tweets                 0
dtype: int64
```

In [11]:

```
df = df.drop('Source of Tweet', axis = 1)
```

In [12]:

```
df
```

Out[12]:

	Date Created	Number of Likes	Tweets
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟\n@go...
2	2023-06-30 09:20:22+00:00	0	@ponilemova #Adipurush
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...
4	2023-06-30 09:15:22+00:00	3	ST: #Adipurush https://t.co/lsGKcgQuKL
...	...	...	...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...
9997	2023-06-23 10:08:49+00:00	0	#GodMorningFriday\nवास्तव में #Adipurush यानि ...
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...

10000 rows × 3 columns

In [13]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10000 entries, 0 to 10000
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Date Created     10000 non-null  object
 1   Number of Likes  10000 non-null  int64 
 2   Tweets           10000 non-null  object
dtypes: int64(1), object(2)
memory usage: 312.5+ KB
```

In [14]:

```
df.describe()
```

Out[14]:

	Number of Likes
count	10000.000000
mean	73.544500
std	369.705682
min	0.000000
25%	0.000000
50%	1.000000
75%	9.000000
max	14778.000000

In [15]:

```
df.nunique()
```

Out[15]:

```
Date Created       9831
Number of Likes     718
Tweets             9874
dtype: int64
```

In [16]:

```
df_sorted = df.sort_values(by='Number of Likes', ascending=False)
```

In [17]:

```
df_sorted.head(10)
```

Out[17]:

	Date Created	Number of Likes	Tweets
5036	2023-06-26 02:51:52+00:00	14778	Pan India Star #Prabhas clearly said NO for #A...
2975	2023-06-27 12:35:31+00:00	8266	#Breaking: Comments by Allahabad high court to...
8180	2023-06-24 09:10:09+00:00	8112	#Adipurush #Prabhas #BhushanKumar https://t.co...
3593	2023-06-27 02:23:37+00:00	7010	आदिपुरुष निर्माताओं को लगा एक और झटका, इलाहाबा...
6069	2023-06-25 07:20:59+00:00	5580	👉#AdiPurush Telugu Version Hits 100CR SHARE 💥 💥 💥 ...
3601	2023-06-27 01:59:03+00:00	5149	#Adipurush WW BO\n\nZOOMS past ₹4□0□0□ cr.\n\n...
4744	2023-06-26 06:20:58+00:00	4912	#Adipurush goes from strength to strength at t...
4716	2023-06-26 06:30:01+00:00	4788	We are incredibly touched by the overwhelming ...
5636	2023-06-25 13:08:17+00:00	4741	Witness the epic saga unfold! 🏹 \nBook your tic...
1559	2023-06-28 14:40:47+00:00	4561	कुरान पर गलत तथ्यों के साथ एक छोटी सी डॉक्यूमे...

In [18]:

```
df['Date Created'] = pd.to_datetime(df['Date Created'])
```

In [19]:

```
df
```

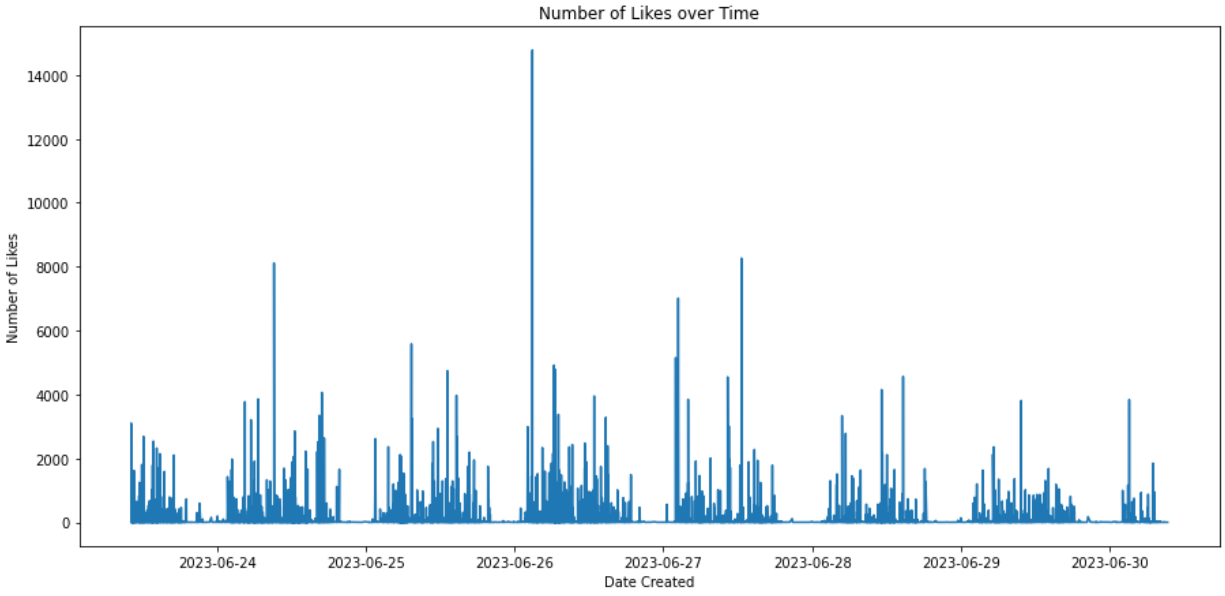
Out[19]:

	Date Created	Number of Likes	Tweets
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟️\n@go...
2	2023-06-30 09:20:22+00:00	0	@ponilemova #Adipurush
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...
4	2023-06-30 09:15:22+00:00	3	ST: #Adipurush https://t.co/lsGKcgQuKL
...	...	...	...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...
9997	2023-06-23 10:08:49+00:00	0	#GodMorningFriday\nवास्तव में #Adipurush यानि ...
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...

10000 rows × 3 columns

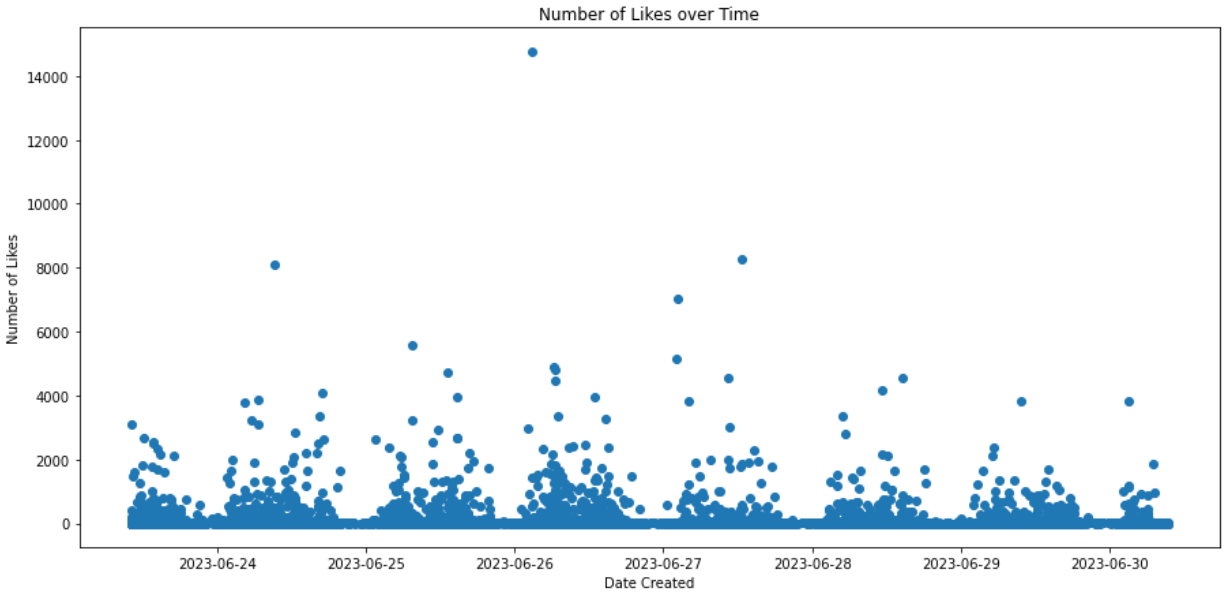
In [20]:

```
df_sorted_date = df.sort_values('Date Created')
plt.figure(figsize=[15,7],)
plt.plot(df_sorted_date['Date Created'], df_sorted_date['Number of Likes'])
plt.xlabel('Date Created')
plt.ylabel('Number of Likes')
plt.title('Number of Likes over Time')
plt.show()
```



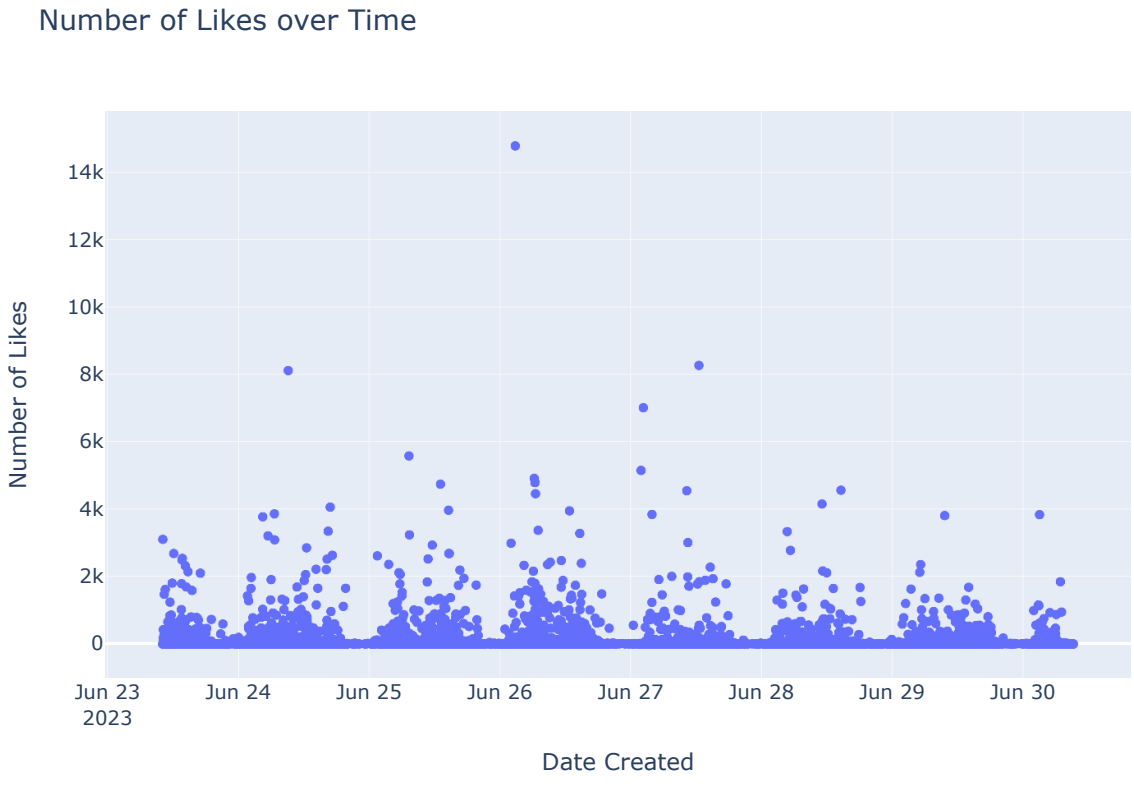
In [21]:

```
plt.figure(figsize=[15,7],)
plt.scatter(df_sorted_date['Date Created'], df_sorted_date['Number of Likes'])
plt.xlabel('Date Created')
plt.ylabel('Number of Likes')
plt.title('Number of Likes over Time')
plt.show()
```



In [22]:

```
import plotly.express as px

fig = px.scatter(df_sorted_date, x='Date Created', y='Number of Likes', title='Number of Likes over Time')
fig.update_layout(xaxis=dict(title='Date Created'), yaxis=dict(title='Number of Likes'))
fig.show()
```



In [23]:

```
import re
import string
from tqdm.notebook import tqdm
from datetime import datetime
import dateutil.parser
```

In [24]:

```
import nltk
from spellchecker import SpellChecker
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
```

In [25]:

```
from wordcloud import WordCloud, ImageColorGenerator
from nltk.corpus import stopwords
import random
```

In [26]:

```
nltk.download('vader_lexicon')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\hp5cd\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\hp5cd\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[26]:

```
True
```

In [27]:

```
languages = stopwords.fileids()

# Print the number of supported languages
print("Number of supported languages:", len(languages))

# Print the list of supported languages
print("Supported languages:", languages)
```

```
Number of supported languages: 29
Supported languages: ['arabic', 'azerbaijani', 'basque', 'bengali', 'catalan', 'chinese', 'danis
h', 'dutch', 'english', 'finnish', 'french', 'german', 'greek', 'hebrew', 'hinglish', 'hungaria
n', 'indonesian', 'italian', 'kazakh', 'nepali', 'norwegian', 'portuguese', 'romanian', 'russia
n', 'slovene', 'spanish', 'swedish', 'tajik', 'turkish']
```

In [28]:

```
from nltk.tokenize import TweetTokenizer
```

In [29]:

```
english_stopwords = stopwords.words('english')
hinglish_stopwords = stopwords.words('hinglish')
```

In [30]:

```
def clean_tweet(tweet):
    # Remove URLs, hashtags, mentions, and special characters
    tweet = re.sub(r"http\S+|www\S+|@\w+|#\w+", "", tweet)
    tweet = re.sub(r"[^\w\s]", "", tweet)

    # Tokenize the tweet
    tokenizer = TweetTokenizer(preserve_case=False, reduce_len=True, strip_handles=True)
    tokens = tokenizer.tokenize(tweet)

    # Remove stopwords for English and Hinglish
    tokens = [token for token in tokens if token not in english_stopwords and token not in hinglish_stopwords]

    # Remove punctuation and convert to lowercase
    tokens = [token.translate(str.maketrans('', '', string.punctuation)) for token in tokens]
    tokens = [token.lower() for token in tokens]

    # Join tokens back into a string
    cleaned_tweet = ' '.join(tokens)

    return cleaned_tweet
```

In [31]:

```
df['Cleaned_Tweets'] = df['Tweets'].apply(clean_tweet)
```

In [32]:

```
df
```

Out[32]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes 2023 live streaming broadcast tv ...
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟️\n@go...	playing book ticket
2	2023-06-30 09:20:22+00:00	0	@ponilemova #Adipurush	
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs 72 hoorain vs kerala story contro...
4	2023-06-30 09:15:22+00:00	3	ST: #Adipurush https://t.co/lsGKcgQuKL	st
...	...	...	...	...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...	adipurush 1st week ww box office collections ఆ...
9997	2023-06-23 10:08:49+00:00	0	#GodMorningFriday\nवास्तव में #Adipurush यानि ...	वसतव म यन सबस पहल भगवन जसन सरव सषट क रचन क ह व...
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...

10000 rows × 4 columns

In [33]:

```
def clean_text(text):
    text = text.lower()
    return text.strip()
```

In [34]:

```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: clean_text(x))
```

In [35]:

```
def tokenization(text):
    tokens = re.split('W+',text)
    return tokens
```

In [36]:

```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: tokenization(x))
```

In [37]:

```
from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()
```

In [38]:

```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\hp5cd\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[38]:

```
True
```

In [39]:

```
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\hp5cd\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

Out[39]:

```
True
```

In [40]:

```
def lemmatizer(text):
    lemm_text = "".join([wordnet_lemmatizer.lemmatize(word) for word in text])
    return lemm_text
```

In [41]:

```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: lemmatizer(x))
```

In [42]:

```
def remove_digits(text):
    clean_text = re.sub(r"\b[0-9]+\b\s*", "", text)
    return(text)
```

In [43]:

```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: remove_digits(x))
```

In [44]:

```
def remove_digits1(sample_text):
    clean_text = " ".join([w for w in sample_text.split() if not w.isdigit()])
    return(clean_text)
```

In [45]:

```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: remove_digits1(x))
```

In [46]:

```
from langdetect import detect

def detect_language(text):
    try:
        lang = detect(text)
        return lang
    except:
        return None

df['Language'] = df['Cleaned_Tweets'].apply(detect_language)
```

In [47]:

```
df
```

Out[47]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥🍿🎟\n@go...	playing book ticket	en
2	2023-06-30 09:20:22+00:00	0	@ponilemova #Adipurush		None
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs hoorain vs kerala story controver...	en
4	2023-06-30 09:15:22+00:00	3	ST: #Adipurush https://t.co/IsGKcgQuKL	st	sv
...	...	...	...	...	...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...	adipurush 1st week ww box office collections ఆ...	te
9997	2023-06-23 10:08:49+00:00	0	#GodMorningFriday\nवास्तव में #Adipurush यानि ...	वसतव म यन सबस पहल भगवन जसन सरव सषट क रचन क ह व...	hi
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...	en
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...	en

10000 rows × 5 columns

In [48]:

```
df1 = df.copy()
```

In [49]:

```
df1['english_tweets'] = df[df['Language'] == 'en']['Cleaned_Tweets']
```

In [50]:

```
df1
```

Out[50]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language	english_tweets
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en	womens ashes live streaming broadcast tv chann...
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟 \n@go...	playing book ticket	en	playing book ticket
2	2023-06-30 09:20:22+00:00	0	@ponilemova #Adipurush		None	NaN
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs hoorain vs kerala story controver...	en	adipurush vs hoorain vs kerala story controver...
4	2023-06-30 09:15:22+00:00	3	ST: #Adipurush https://t.co/lsGKcgQuKL	st	sv	NaN
...	...	...	...	...	...	...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...	adipurush 1st week ww box office collections ఆ...	te	NaN
9997	2023-06-23 10:08:49+00:00	0	#GodMorningFriday\nवास्तव में #Adipurush यानि ...	वसतव म यन सबस पहल भगवन जसन सरव सषट क रचन क ह व...	hi	NaN
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...	en	choosing service product beneficial opt authen...
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...	en	film greatest epic earn boc worth budget shame...

10000 rows × 6 columns

In [51]:

```
df1 = df1.dropna()
```

In [52]:

```
df1
```

Out[52]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language	english_tweets
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en	womens ashes live streaming broadcast tv chann...
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟️ \n@go...	playing book ticket	en	playing book ticket
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs hoorain vs kerala story controver...	en	adipurush vs hoorain vs kerala story controver...
5	2023-06-30 09:08:27+00:00	1	This is how the story should be told. @omraut ...	story told learn hotstar india graphic india g...	en	story told learn hotstar india graphic india g...
8	2023-06-30 09:04:09+00:00	0	@VikasAgarwalll Milord says: If my compatriots...	milord compatriots backstab ie end exposing fa...	en	milord compatriots backstab ie end exposing fa...
...	...	...	...	...	...	...
9994	2023-06-23 10:10:47+00:00	1	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...
9995	2023-06-23 10:09:41+00:00	1	S Rangarajan garu, main poojari of chilkur bal...	rangarajan garu poojari chilkur balaji appreci...	en	rangarajan garu poojari chilkur balaji appreci...
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...	en	choosing service product beneficial opt authen...
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...	en	film greatest epic earn boc worth budget shame...

5105 rows × 6 columns

In [53]:

```
df1['Year'] = df1['Date Created'].dt.year
df1['Month'] = df1['Date Created'].dt.month
df1['Day'] = df1['Date Created'].dt.day
```

In [54]:

```
df1
```

Out[54]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language	english_tweets	Year	Month	Day
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en	womens ashes live streaming broadcast tv chann...	2023	6	30
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟️\n@go...	playing book ticket	en	playing book ticket	2023	6	30
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs hoorain vs kerala story controver...	en	adipurush vs hoorain vs kerala story controver...	2023	6	30
5	2023-06-30 09:08:27+00:00	1	This is how the story should be told. @omraut ...	story told learn hotstar india graphic india g...	en	story told learn hotstar india graphic india g...	2023	6	30
8	2023-06-30 09:04:09+00:00	0	@VikasAgarwalll Milord says: If my compatriots...	milord compatriots backstab ie end exposing fa...	en	milord compatriots backstab ie end exposing fa...	2023	6	30
...	...	...	...	...	...	...	...	...	...
9994	2023-06-23 10:10:47+00:00	1	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...	2023	6	23
9995	2023-06-23 10:09:41+00:00	1	S Rangarajan garu, main poojari of chilkur bal...	rangarajan garu poojari chilkur balaji appreci...	en	rangarajan garu poojari chilkur balaji appreci...	2023	6	23
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...	2023	6	23
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...	en	choosing service product beneficial opt authen...	2023	6	23
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...	en	film greatest epic earn boc worth budget shame...	2023	6	23

5105 rows × 9 columns

In [55]:

```
df1.nunique()
```

Out[55]:

```
Date Created       5067
Number of Likes     565
Tweets             5058
Cleaned_Tweets     4471
Language              1
english_tweets     4471
Year                  1
Month                 1
Day                   8
dtype: int64
```

In [56]:

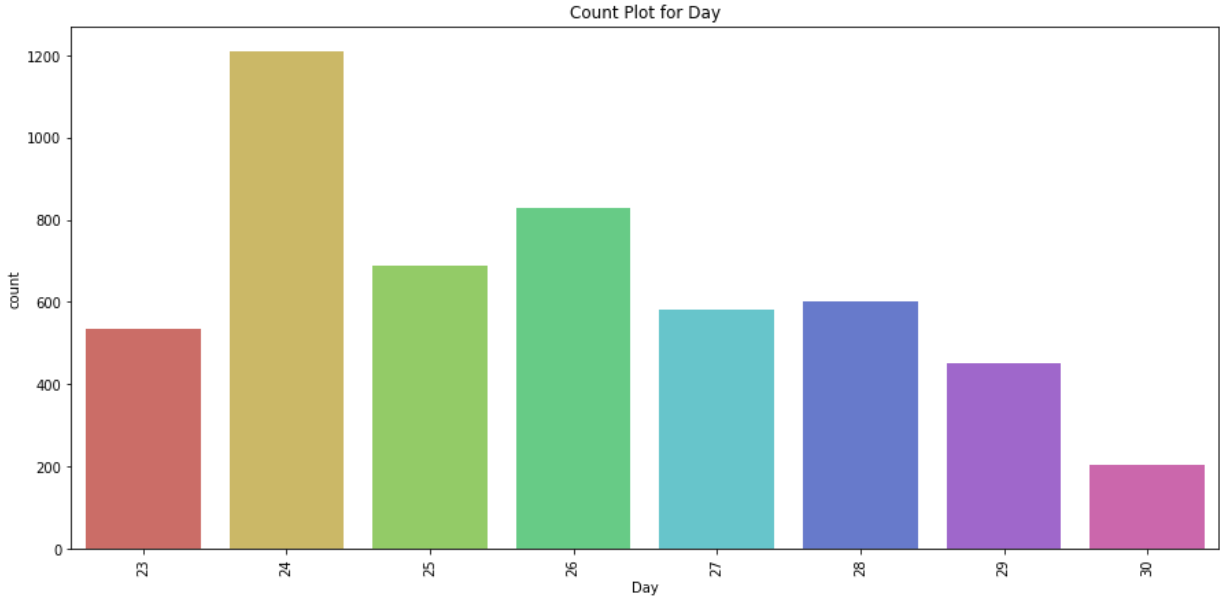
```
df1['Time'] = df1['Date Created'].dt.time
```

In [57]:

```
df1['Tweet_Length'] = df1['english_tweets'].str.len()
```

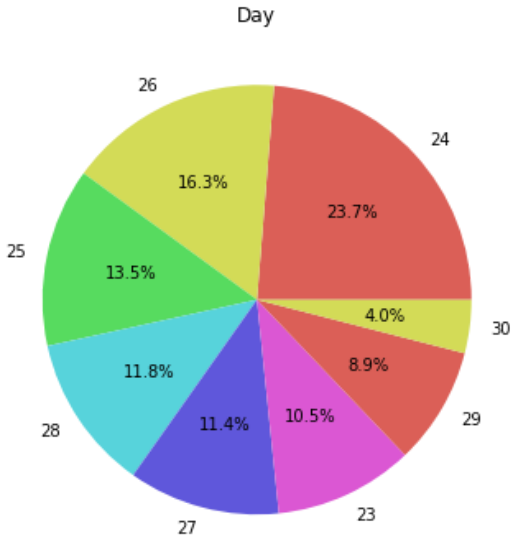
In [58]:

```
plt.figure(figsize=[15,7],)
plt.title('Count Plot for Day')
sns.countplot(x = 'Day', data = df1, palette = 'hls')
plt.xticks(rotation = 90)
plt.show()
```



In [59]:

```
plt.figure(figsize=(15, 6))
counts = df1['Day'].value_counts()
plt.pie(counts, labels=counts.index, autopct='%1.1f%%', colors=sns.color_palette('hls'))
plt.title('Day')
plt.show()
```

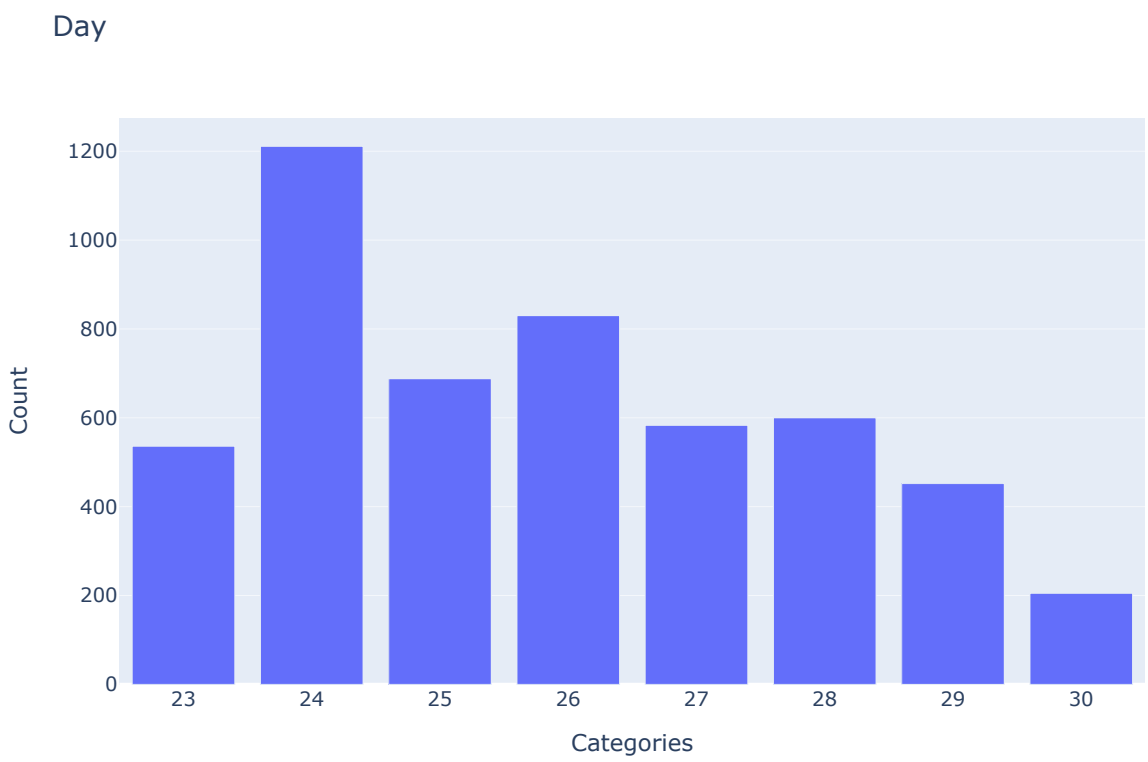


In [60]:

```
import plotly.graph_objects as go
```

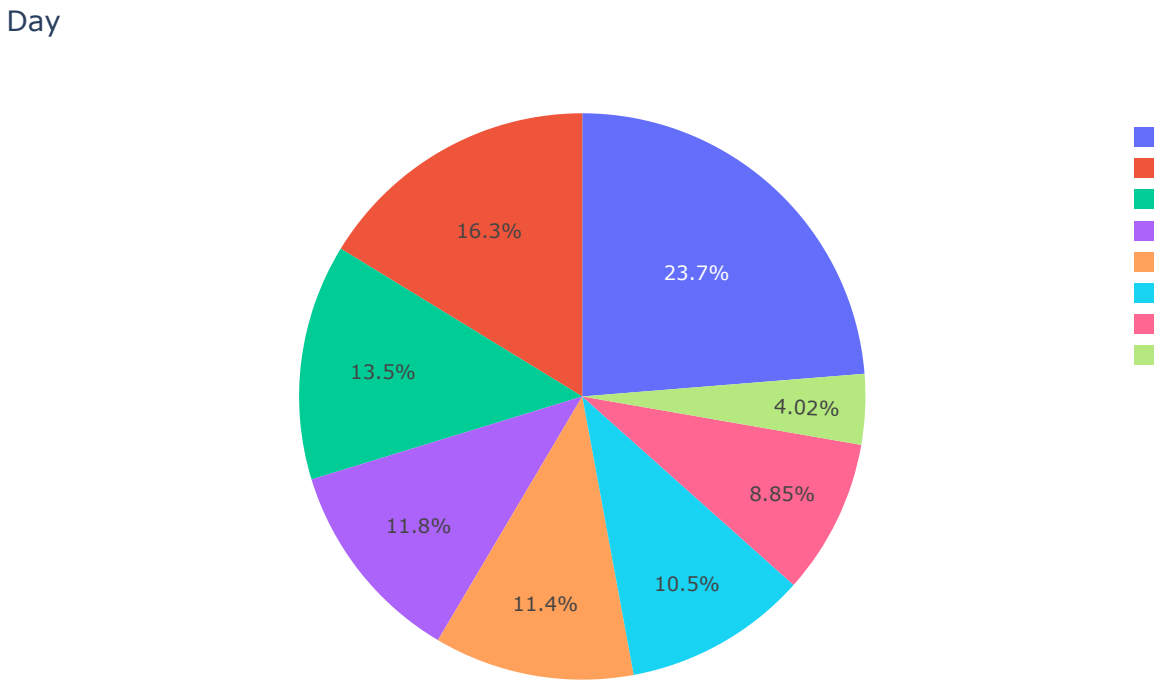
In [61]:

```
fig = go.Figure(data=[go.Bar(x=df1['Day'].value_counts().index, y=df1['Day'].value_counts())])
fig.update_layout(
        title= 'Day',
        xaxis_title="Categories",
        yaxis_title="Count"
    )
fig.show()
```



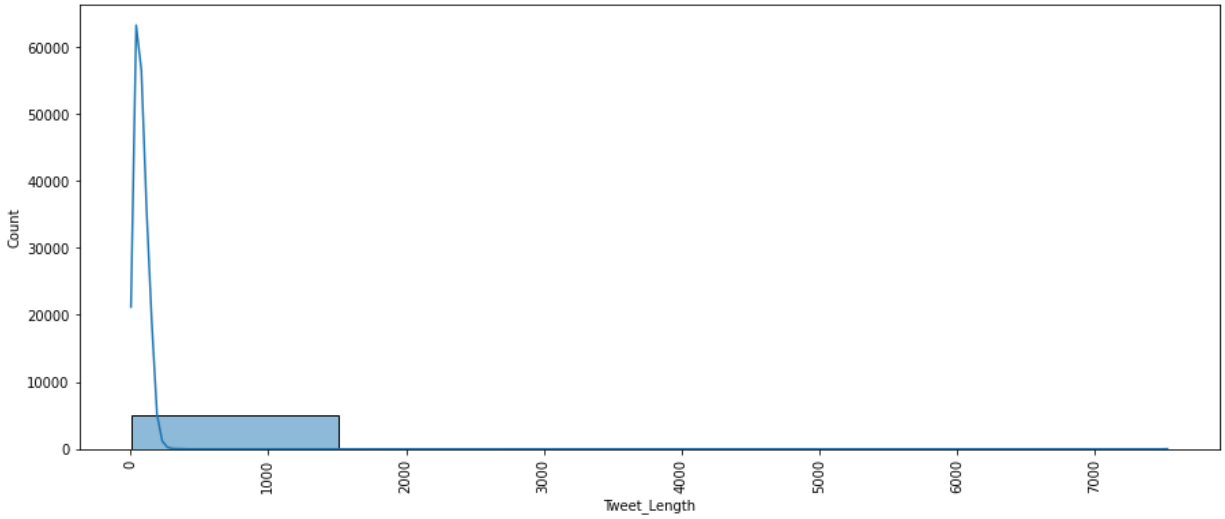
In [62]:

```
counts = df1['Day'].value_counts()
fig = go.Figure(data=[go.Pie(labels=counts.index, values=counts)])
fig.update_layout(title= 'Day')
fig.show()
```



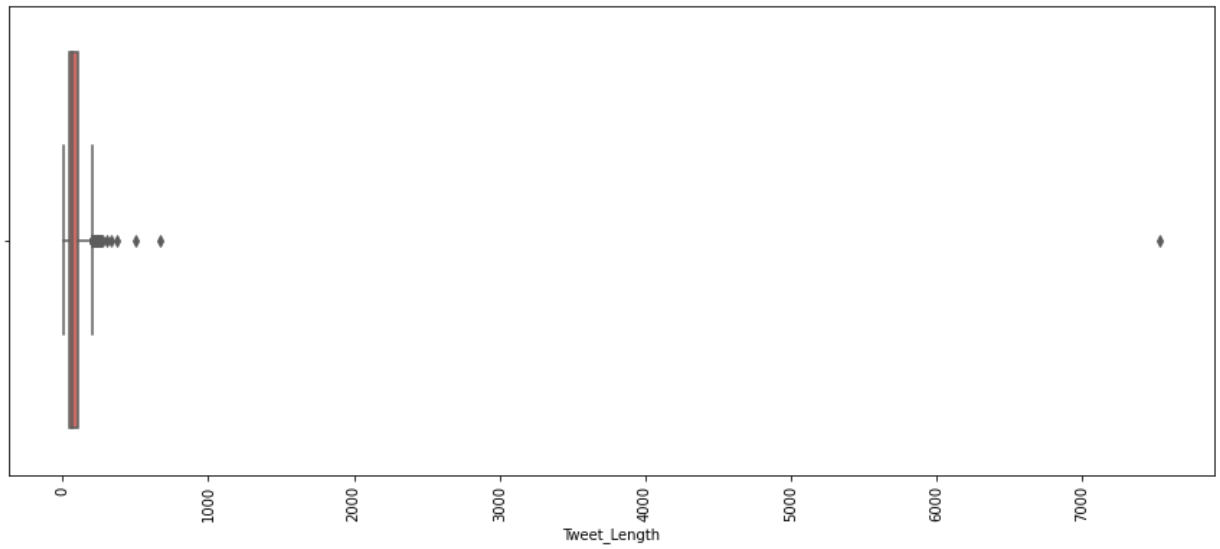
In [63]:

```
plt.figure(figsize=(15,6))
sns.histplot(df1['Tweet_Length'], kde = True, bins = 5, palette = 'hls')
plt.xticks(rotation = 90)
plt.show()
```



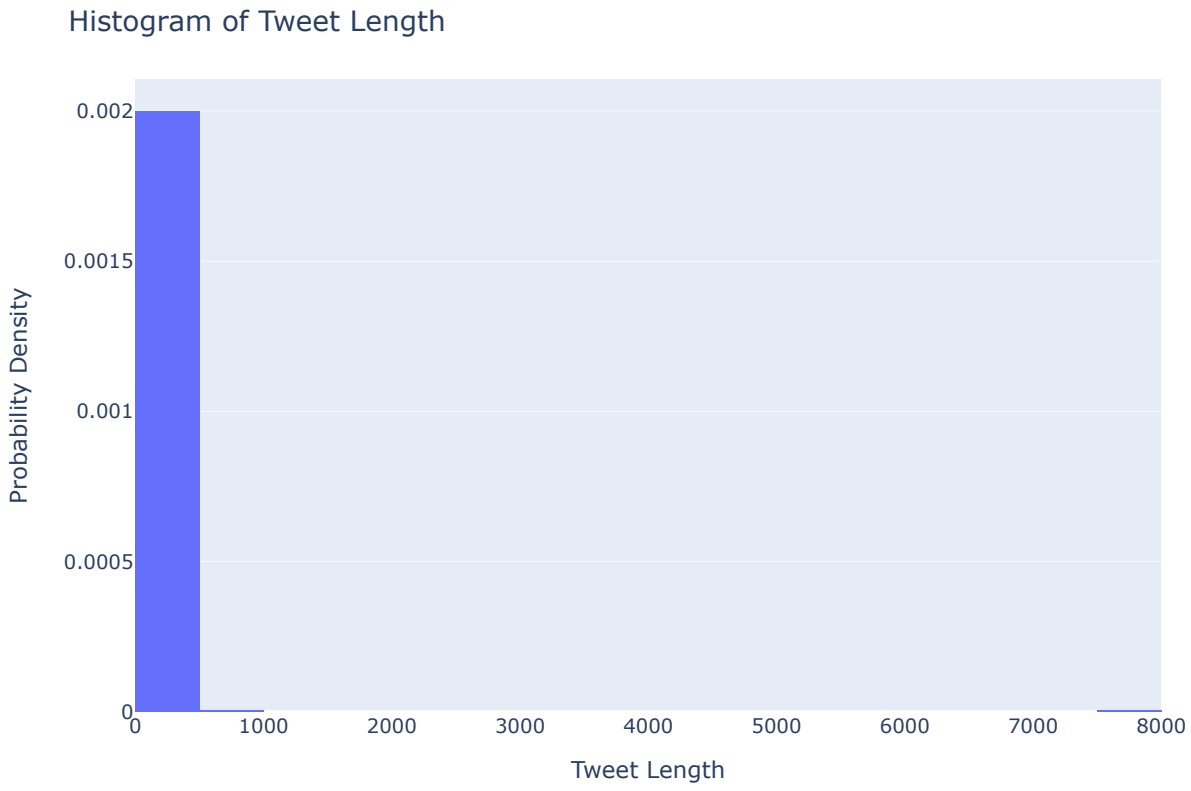
In [64]:

```
plt.figure(figsize=(15,6))
sns.boxplot(df1['Tweet_Length'], data = df, palette = 'hls')
plt.xticks(rotation = 90)
plt.show()
```



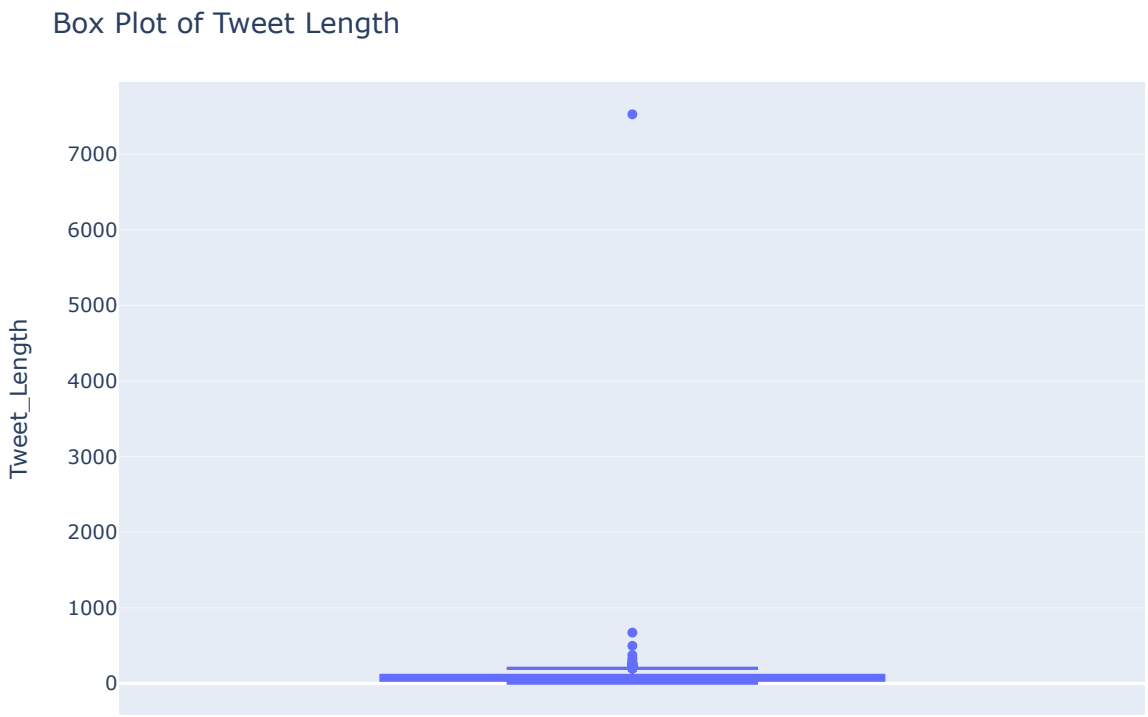
In [65]:

```
import plotly.express as px

fig = px.histogram(df1, x='Tweet_Length', nbins=20, histnorm='probability density')
fig.update_layout(title=f"Histogram of Tweet Length", xaxis_title='Tweet Length', yaxis_title="Probability Dens
fig.show()
```



In [66]:

```
fig = px.box(df1, y='Tweet_Length')
fig.update_layout(title=f"Box Plot of Tweet Length", yaxis_title='Tweet_Length')
fig.show()
```



In [67]:

```
spell = SpellChecker()
```

In [68]:

```
def label_sentiment(x:float):
    if x < -0.05 : return 'negative'
    if x > 0.35 : return 'positive'
    return 'neutral'
```

In [69]:

```
sia = SIA()
```

In [70]:

```
df1['sentiment'] = [sia.polarity_scores(x)['compound'] for x in tqdm(df1['english_tweets'])]
df1['overall_sentiment'] = df1['sentiment'].apply(label_sentiment);
```

```
  0%|          | 0/5105 [00:00<?, ?it/s]
```

In [71]:

```
df1
```

Out[71]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language	english_tweets	Year	Month	Day	Time
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en	womens ashes live streaming broadcast tv chann...	2023	6	30	09:21:00
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟 \n@go...	playing book ticket	en	playing book ticket	2023	6	30	09:20:57
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs hoorain vs kerala story controver...	en	adipurush vs hoorain vs kerala story controver...	2023	6	30	09:20:00
5	2023-06-30 09:08:27+00:00	1	This is how the story should be told. @omraut ...	story told learn hotstar india graphic india g...	en	story told learn hotstar india graphic india g...	2023	6	30	09:08:27
8	2023-06-30 09:04:09+00:00	0	@VikasAgarwalll Milord says: If my compatriots...	milord compatriots backstab ie end exposing fa...	en	milord compatriots backstab ie end exposing fa...	2023	6	30	09:04:09
...	...	...	...	...	...	...	...	...	...	...
9994	2023-06-23 10:10:47+00:00	1	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...	2023	6	23	10:10:47
9995	2023-06-23 10:09:41+00:00	1	S Rangarajan garu, main poojari of chilkur bal...	rangarajan garu poojari chilkur balaji appreci...	en	rangarajan garu poojari chilkur balaji appreci...	2023	6	23	10:09:41
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...	2023	6	23	10:08:17
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...	en	choosing service product beneficial opt authen...	2023	6	23	10:08:01
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...	en	film greatest epic earn boc worth budget shame...	2023	6	23	10:07:45

5105 rows × 13 columns

In [72]:

```
df1['overall_sentiment'].unique()
```

Out[72]:

```
array(['neutral', 'positive', 'negative'], dtype=object)
```

In [73]:

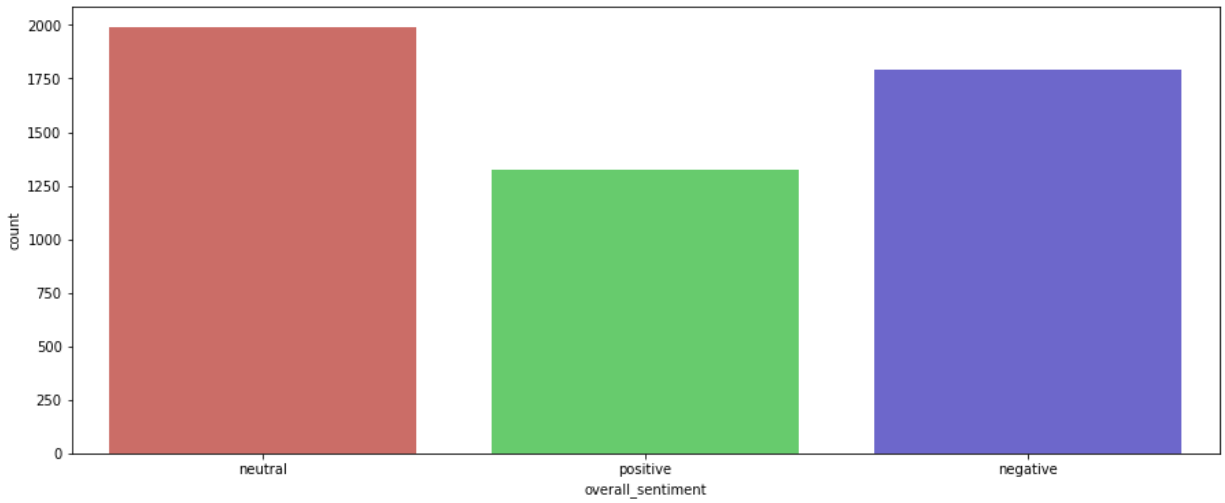
```
df1['overall_sentiment'].value_counts()
```

Out[73]:

```
neutral     1988
negative    1794
positive    1323
Name: overall_sentiment, dtype: int64
```

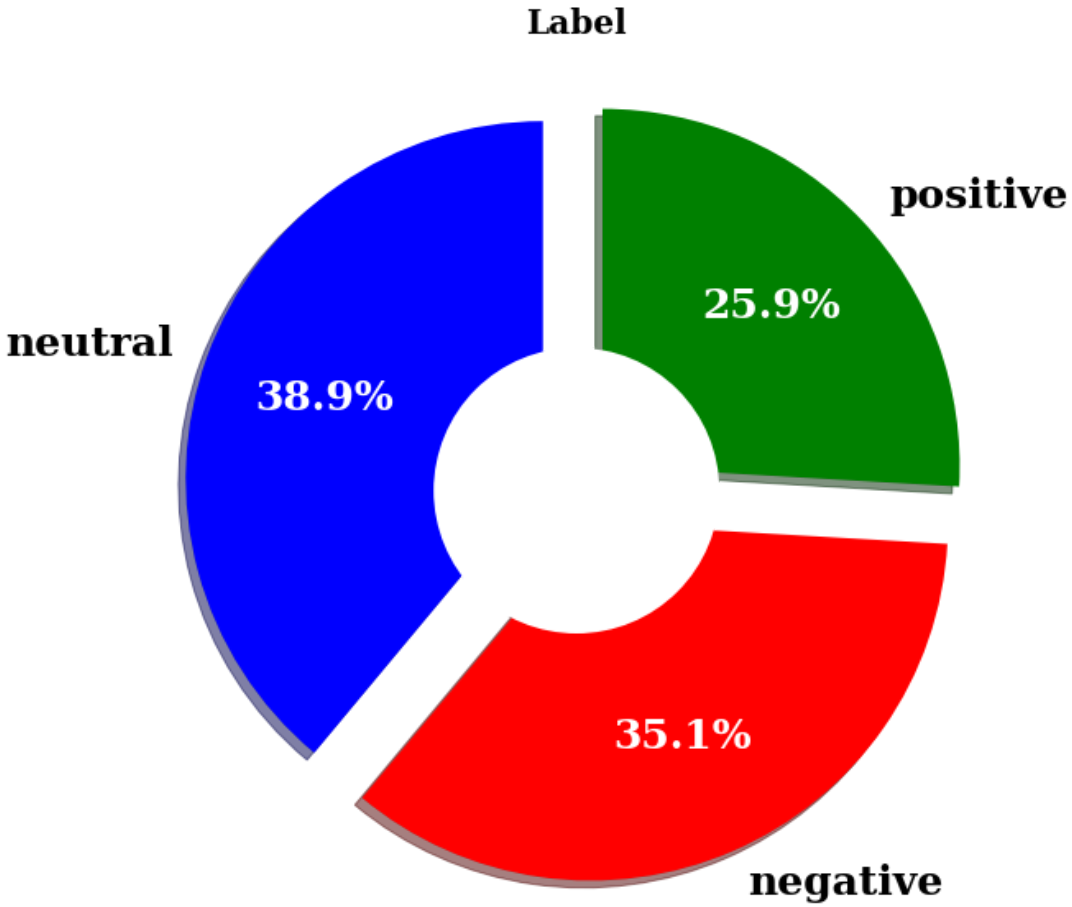
In [74]:

```
plt.figure(figsize=(15,6))
sns.countplot(df1['overall_sentiment'], data = df1, palette = 'hls')
plt.xticks(rotation = 0)
plt.show()
```



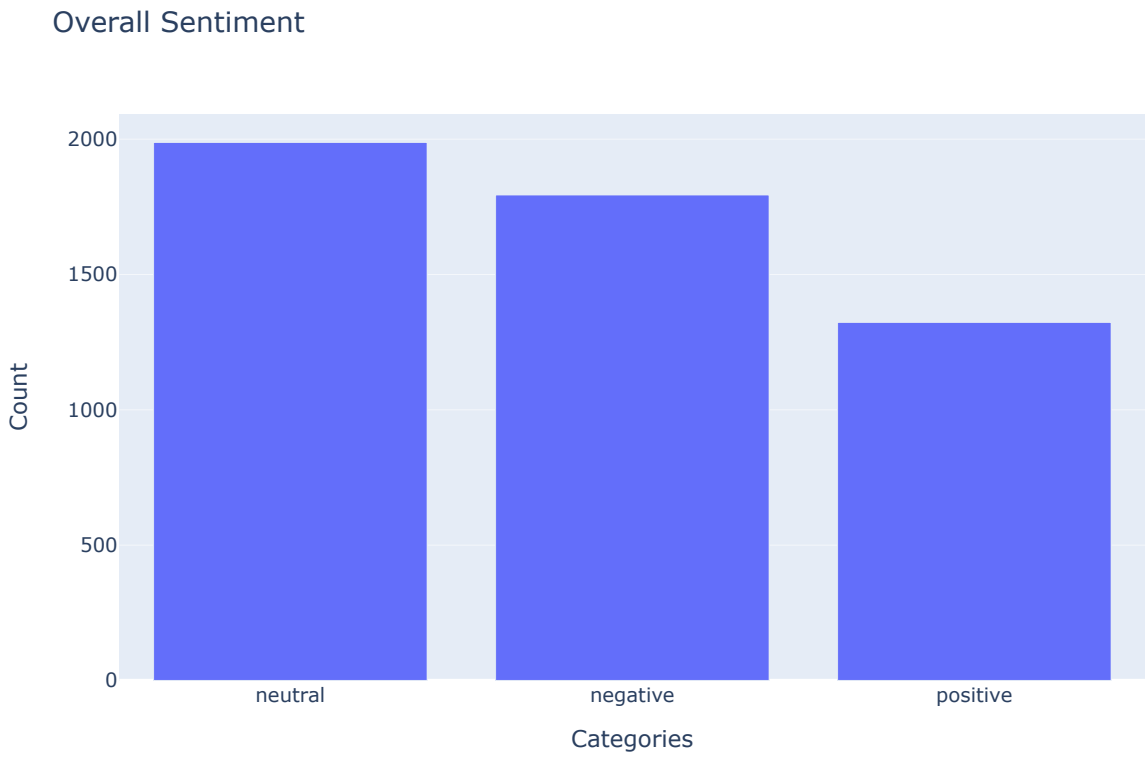
In [75]:

```
label_data = df1['overall_sentiment'].value_counts()

explode = (0.1, 0.1, 0.1)
plt.figure(figsize=(14, 10))
patches, texts, pcts = plt.pie(label_data,
                               labels = label_data.index,
                               colors = ['blue', 'red', 'green'],
                               pctdistance = 0.65,
                               shadow = True,
                               startangle = 90,
                               explode = explode,
                               autopct = '%1.1f%%',
                               textprops={ 'fontsize': 25,
                                           'color': 'black',
                                           'weight': 'bold',
                                           'family': 'serif' })
plt.setp(pcts, color='white')

hfont = {'fontname':'serif', 'weight': 'bold'}
plt.title('Label', size=20, **hfont)

centre_circle = plt.Circle((0,0),0.40,fc='white')
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.show()
```



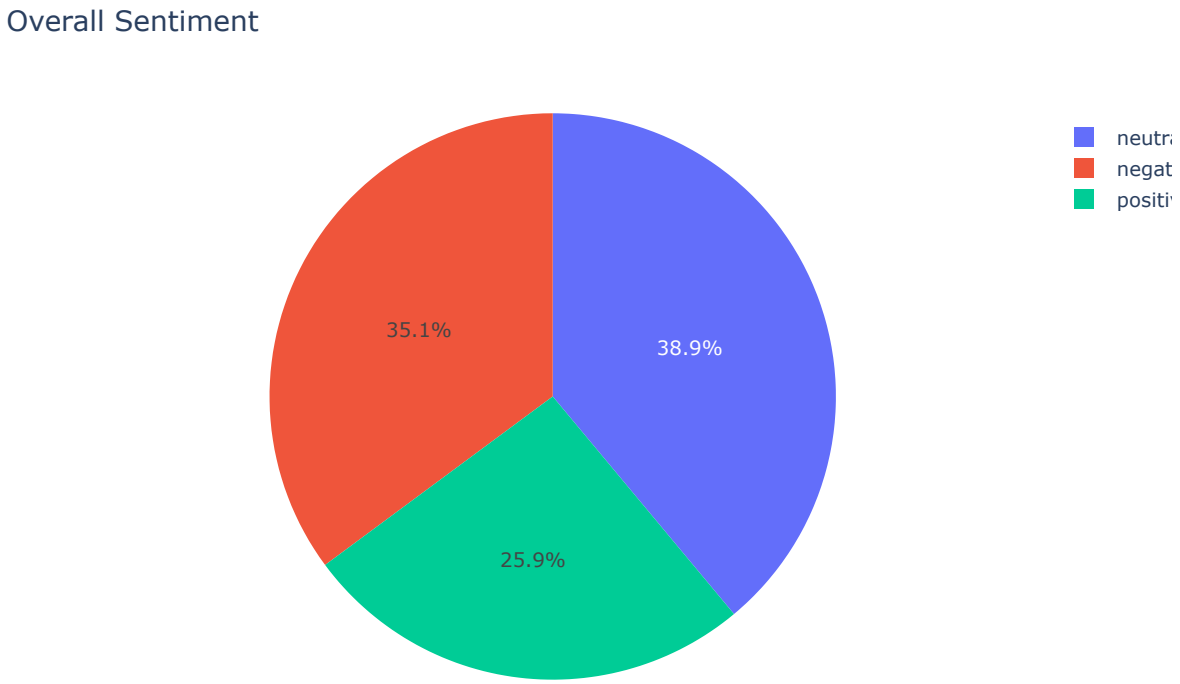
In [76]:

```
fig = go.Figure(data=[go.Bar(x=df1['overall_sentiment'].value_counts().index, y=df1['overall_sentiment'].value_
fig.update_layout(
        title= 'Overall Sentiment',
        xaxis_title="Categories",
        yaxis_title="Count"
    )
fig.show()
```



In [77]:

```
counts = df1['overall_sentiment'].value_counts()
fig = go.Figure(data=[go.Pie(labels=counts.index, values=counts)])
fig.update_layout(title= 'Overall Sentiment')
fig.show()
```



In [78]:

```
df1
```

Out[78]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language	english_tweets	Year	Month	Day	Time
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en	womens ashes live streaming broadcast tv chann...	2023	6	30	09:21:00
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥 🍿 🎟 \n@go...	playing book ticket	en	playing book ticket	2023	6	30	09:20:57
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain VS The Kerala Story Co...	adipurush vs hoorain vs kerala story controver...	en	adipurush vs hoorain vs kerala story controver...	2023	6	30	09:20:00
5	2023-06-30 09:08:27+00:00	1	This is how the story should be told. @omraut ...	story told learn hotstar india graphic india g...	en	story told learn hotstar india graphic india g...	2023	6	30	09:08:27
8	2023-06-30 09:04:09+00:00	0	@VikasAgarwalll Milord says: If my compatriots...	milord compatriots backstab ie end exposing fa...	en	milord compatriots backstab ie end exposing fa...	2023	6	30	09:04:09
...	...	...	...	...	...	...	...	...	...	...
9994	2023-06-23 10:10:47+00:00	1	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...	2023	6	23	10:10:47
9995	2023-06-23 10:09:41+00:00	1	S Rangarajan garu, main poojari of chilkur bal...	rangarajan garu poojari chilkur balaji appreci...	en	rangarajan garu poojari chilkur balaji appreci...	2023	6	23	10:09:41
9998	2023-06-23 10:08:17+00:00	3101	Let the empowering lyrics of #Shivoham elevate...	empowering lyrics elevate spirit envelop world...	en	empowering lyrics elevate spirit envelop world...	2023	6	23	10:08:17
9999	2023-06-23 10:08:01+00:00	0	When it comes to choosing a service or product...	choosing service product beneficial opt authen...	en	choosing service product beneficial opt authen...	2023	6	23	10:08:01
10000	2023-06-23 10:07:45+00:00	0	A film about #Ramayana, our greatest epic coul...	film greatest epic earn boc worth budget shame...	en	film greatest epic earn boc worth budget shame...	2023	6	23	10:07:45

5105 rows × 13 columns

In [79]:

```
df2 = df1[['english_tweets', 'overall_sentiment']]
```

In [80]:

```
df2
```

Out[80]:

	english_tweets	overall_sentiment
0	womens ashes live streaming broadcast tv chann...	neutral
1	playing book ticket	neutral
3	adipurush vs hoorain vs kerala story controver...	neutral
5	story told learn hotstar india graphic india g...	neutral
8	milord compatriots backstab ie end exposing fa...	positive
...	...	...
9994	empowering lyrics elevate spirit envelop world...	positive
9995	rangarajan garu poojari chilkur balaji appreci...	positive
9998	empowering lyrics elevate spirit envelop world...	positive
9999	choosing service product beneficial opt authen...	positive
10000	film greatest epic earn boc worth budget shame...	positive

5105 rows × 2 columns

In [81]:

```
def clean_text(text):
    # Remove non-alphabetic characters and convert to lowercase
    cleaned_text = re.sub('[^a-zA-Z]', ' ', text).lower()
    # Remove extra white spaces
    cleaned_text = re.sub('\s+', ' ', cleaned_text).strip()
    # Split the text into words
    words = cleaned_text.split()
    # Join the words back into a string
    cleaned_text = ' '.join(words)
    return cleaned_text

# Apply the clean_text function to the 'english_tweets' column
df2['Cleaned_English_Tweets'] = df2['english_tweets'].apply(clean_text)
```

In [82]:

```
df2
```

Out[82]:

	english_tweets	overall_sentiment	Cleaned_English_Tweets
0	womens ashes live streaming broadcast tv chann...	neutral	womens ashes live streaming broadcast tv chann...
1	playing book ticket	neutral	playing book ticket
3	adipurush vs hoorain vs kerala story controver...	neutral	adipurush vs hoorain vs kerala story controver...
5	story told learn hotstar india graphic india g...	neutral	story told learn hotstar india graphic india g...
8	milord compatriots backstab ie end exposing fa...	positive	milord compatriots backstab ie end exposing fa...
...	...	...	...
9994	empowering lyrics elevate spirit envelop world...	positive	empowering lyrics elevate spirit envelop world...
9995	rangarajan garu poojari chilkur balaji appreci...	positive	rangarajan garu poojari chilkur balaji appreci...
9998	empowering lyrics elevate spirit envelop world...	positive	empowering lyrics elevate spirit envelop world...
9999	choosing service product beneficial opt authen...	positive	choosing service product beneficial opt authen...
10000	film greatest epic earn boc worth budget shame...	positive	film greatest epic earn boc worth budget shame...

5105 rows × 3 columns

In [83]:

```
df3 = df2[['Cleaned_English_Tweets', 'overall_sentiment']]
```

In [84]:

```
df3
```

Out[84]:

	Cleaned_English_Tweets	overall_sentiment
0	womens ashes live streaming broadcast tv chann...	neutral
1	playing book ticket	neutral
3	adipurush vs hoorain vs kerala story controver...	neutral
5	story told learn hotstar india graphic india g...	neutral
8	milord compatriots backstab ie end exposing fa...	positive
...	...	...
9994	empowering lyrics elevate spirit envelop world...	positive
9995	rangarajan garu poojari chilkur balaji appreci...	positive
9998	empowering lyrics elevate spirit envelop world...	positive
9999	choosing service product beneficial opt authen...	positive
10000	film greatest epic earn boc worth budget shame...	positive

5105 rows × 2 columns

In [85]:

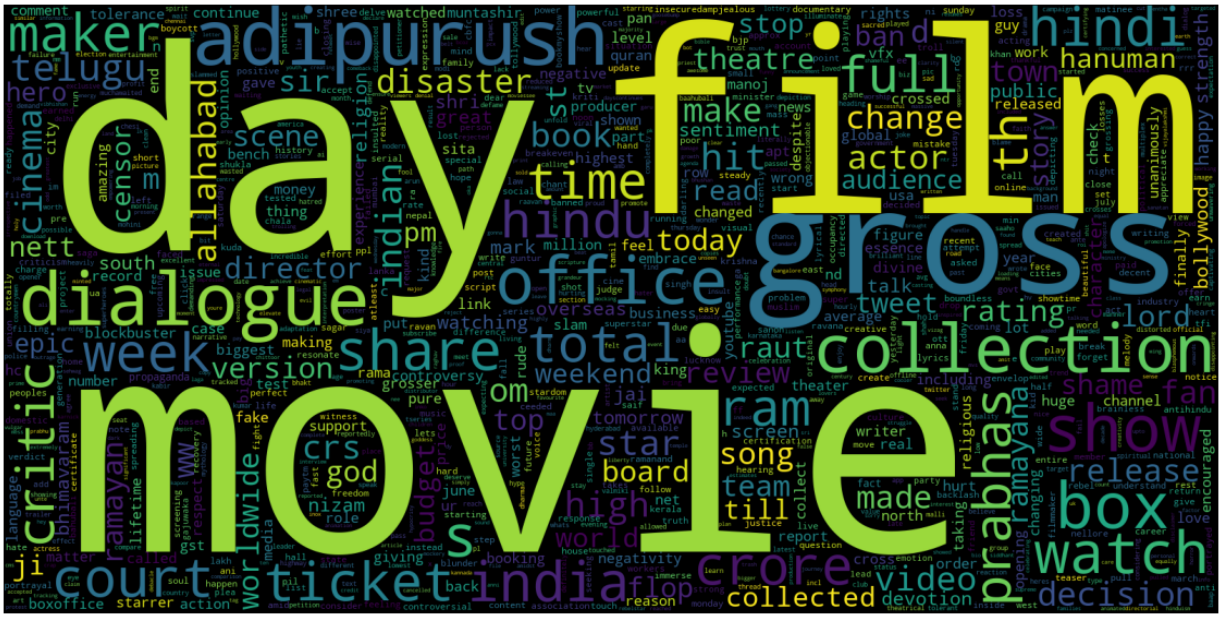
```
non_meaningful_words = ['cr', 'amp', 'rs', 'u', 'l']

def remove_non_meaningful_words(text):
    tokens = text.split()
    filtered_tokens = [token for token in tokens if token not in non_meaningful_words]
    filtered_text = ' '.join(filtered_tokens)
    return filtered_text

df3['Cleaned_English_Tweets'] = df3['Cleaned_English_Tweets'].apply(remove_non_meaningful_words)
```

In [86]:

```
import wordcloud
```

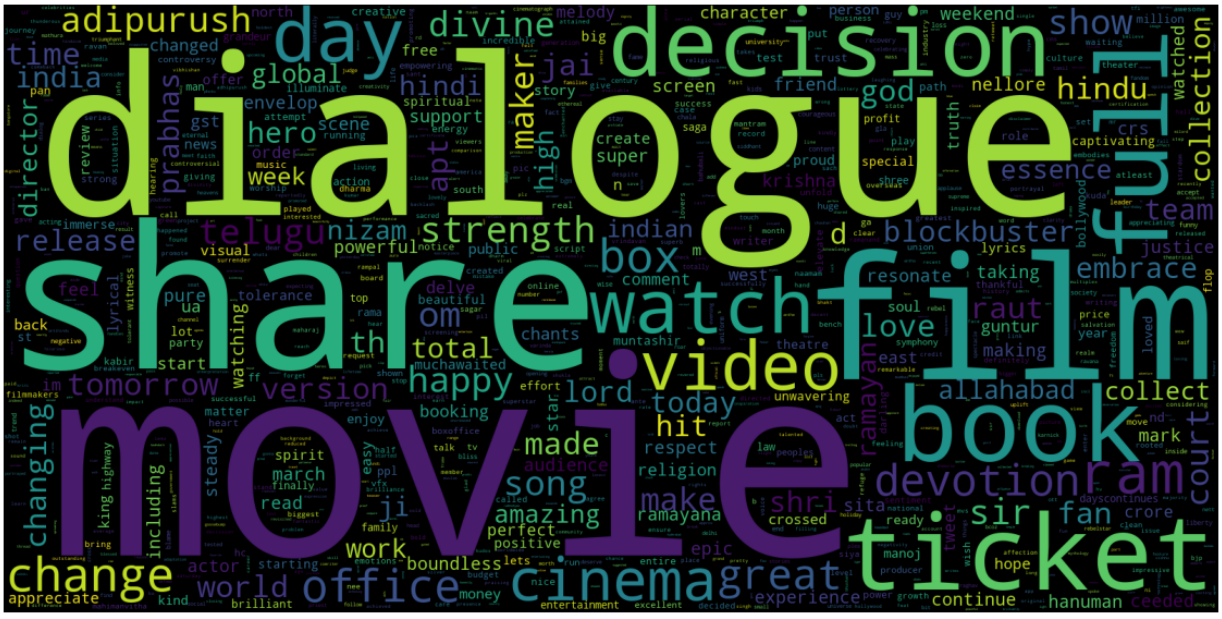
In [87]:

```
from wordcloud import WordCloud
data = df3['Cleaned_English_Tweets']
plt.figure(figsize = (20,20))
wc = WordCloud(max_words = 1000 , width = 1600 , height = 800,
               collocations=False).generate(" ".join(data))
plt.imshow(wc)
plt.axis('off')
plt.show()
```



In [88]:

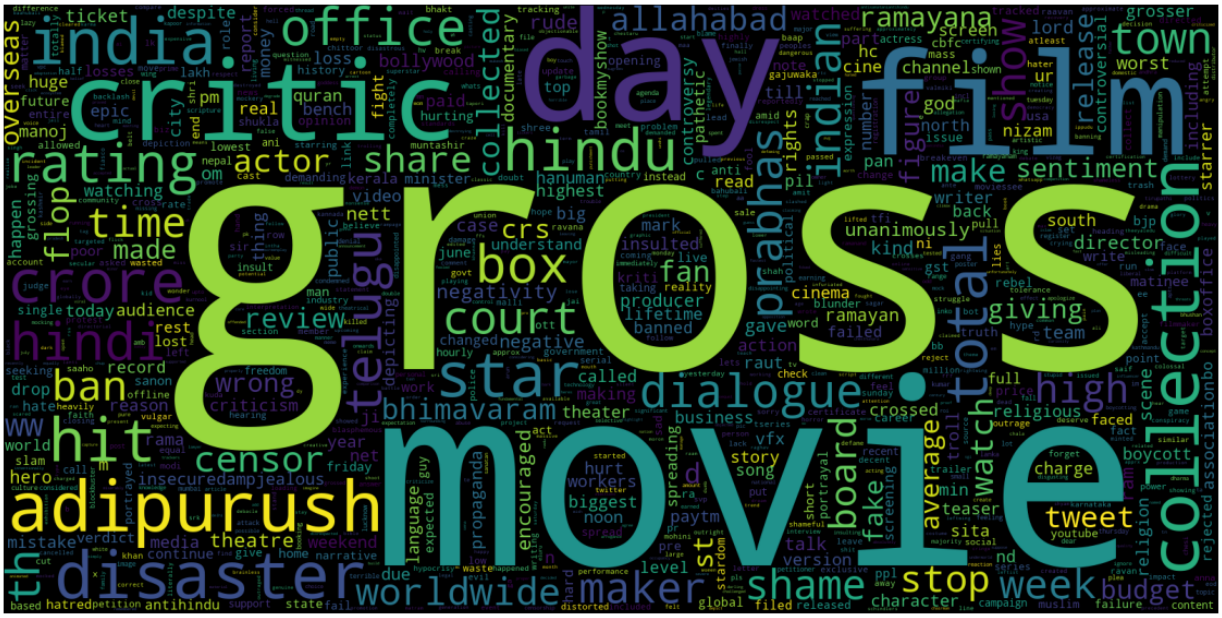
```
data = df3[df3['overall_sentiment']=="positive"]['Cleaned_English_Tweets']
plt.figure(figsize = (20,20))
wc = WordCloud(max_words = 1000 , width = 1600 , height = 800,
               collocations=False).generate(" ".join(data))
plt.imshow(wc)
plt.axis('off')
plt.show()
```



from PIL import Image import numpy as np data = df3[df3['overall_sentiment'] == "positive"]['Cleaned_English_Tweets'] text = " ".join(data) wordcloud = WordCloud(max_words=1000, width=1600, height=800, collocations=False).generate(text) image = wordcloud.to_image() image_array = np.array(image) fig = go.Figure(data=go.Image(z=image_array)) fig.update_layout(title_text="Word Cloud - Positive Tweets", width=800, height=600) fig.show()

In [89]:

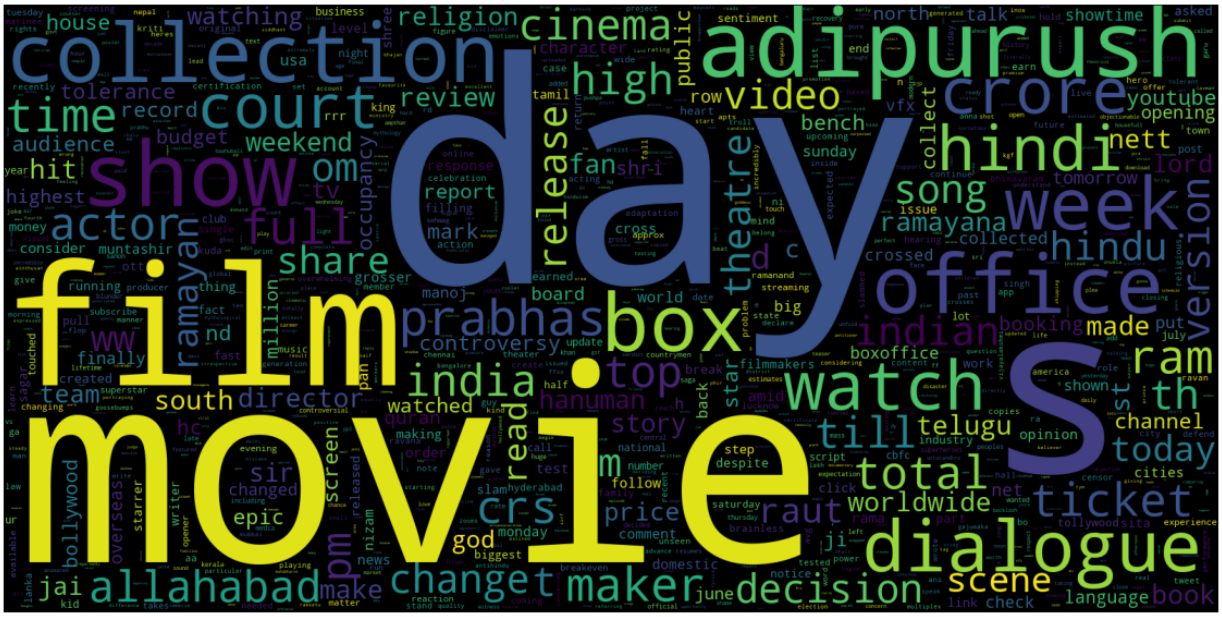
```
data = df3[df3['overall_sentiment']=="negative"]['Cleaned_English_Tweets']
plt.figure(figsize = (20,20))
wc = WordCloud(max_words = 1000 , width = 1600 , height = 800,
               collocations=False).generate(" ".join(data))
plt.imshow(wc)
plt.axis('off')
plt.show()
```



from PIL import Image import numpy as np data = df3[df3['overall_sentiment'] == "negative"]['Cleaned_English_Tweets'] text = " ".join(data) wordcloud = WordCloud(max_words=1000, width=1600, height=800, collocations=False).generate(text) image = wordcloud.to_image() image_array = np.array(image) fig = go.Figure(data=go.Image(z=image_array)) fig.update_layout(title_text="Word Cloud - Negative Tweets", width=800, height=600) fig.show()

In [90]:

```
data = df3[df3['overall_sentiment']=="neutral"]['Cleaned_English_Tweets']
plt.figure(figsize = (20,20))
wc = WordCloud(max_words = 1000 , width = 1600 , height = 800,
               collocations=False).generate(" ".join(data))
plt.imshow(wc)
plt.axis('off')
plt.show()
```



import numpy as np data = df3[df3['overall_sentiment'] == "neutral"]['Cleaned_English_Tweets'] text = " ".join(data) wordcloud = WordCloud(max_words=1000, width=1600, height=800, collocations=False).generate(text) image = wordcloud.to_image() image_array = np.array(image) fig = go.Figure(data=go.Image(z=image_array)) fig.update_layout(title_text="Word Cloud - Neutral Tweets", width=800, height=600) fig.show()

In [91]:

```
df3
```

Out[91]:

	Cleaned_English_Tweets	overall_sentiment
0	womens ashes live streaming broadcast tv chann...	neutral
1	playing book ticket	neutral
3	adipurush vs hoorain vs kerala story controver...	neutral
5	story told learn hotstar india graphic india g...	neutral
8	milord compatriots backstab ie end exposing fa...	positive
...	...	...
9994	empowering lyrics elevate spirit envelop world...	positive
9995	rangarajan garu poojari chilkur balaji appreci...	positive
9998	empowering lyrics elevate spirit envelop world...	positive
9999	choosing service product beneficial opt authen...	positive
10000	film greatest epic earn boc worth budget shame...	positive

5105 rows × 2 columns

In [92]:

```
x = df3['Cleaned_English_Tweets']
y = df3['overall_sentiment']

print(len(x), len(y))
```

```
5105 5105
```

In [93]:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=42)
print(len(x_train), len(y_train))
print(len(x_test), len(y_test))
```

```
3828 3828
1277 1277
```

In [94]:

```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer()
vect.fit(x_train)
```

Out[94]:

```
CountVectorizer
CountVectorizer()
```

In [95]:

```
x_train_dtm = vect.transform(x_train)
x_test_dtm = vect.transform(x_test)
```

In [96]:

```
vect_tunned = CountVectorizer(stop_words='english', ngram_range=(1,2), min_df=0.1, max_df=0.7, max_features=100
```

In [97]:

```
from sklearn.feature_extraction.text import TfidfTransformer

tfidf_transformer = TfidfTransformer()

tfidf_transformer.fit(x_train_dtm)
x_train_tfidf = tfidf_transformer.transform(x_train_dtm)

x_train_tfidf
```

Out[97]:

```
<3828x8210 sparse matrix of type '<class 'numpy.float64'>'
        with 38272 stored elements in Compressed Sparse Row format>
```

In [98]:

```
texts = df3['Cleaned_English_Tweets']
target = df3['overall_sentiment']
```

In [99]:

```
from keras.preprocessing.text import Tokenizer
```

In [100]:

```
word_tokenizer = Tokenizer()
word_tokenizer.fit_on_texts(texts)

vocab_length = len(word_tokenizer.word_index) + 1
vocab_length
```

Out[100]:

```
9638
```

In [101]:

```
import tensorflow as tf
from tensorflow.keras.preprocessing.sequence import pad_sequences
from nltk.tokenize import word_tokenize
```

In [102]:

```
def embed(corpus):
    return word_tokenizer.texts_to_sequences(corpus)

longest_train = max(texts, key=lambda sentence: len(word_tokenize(sentence)))
length_long_sentence = len(word_tokenize(longest_train))

train_padded_sentences = pad_sequences(
    embed(texts),
    length_long_sentence,
    padding='post'
)

train_padded_sentences
```

Out[102]:

```
array([[2272, 1875,  332, ...,    0,    0,    0],
       [ 489,   35,  104, ...,    0,    0,    0],
       [   5,  490, 2275, ...,    0,    0,    0],
       ...,
       [ 870,  439,  812, ...,    0,    0,    0],
       [2757, 1542, 1412, ...,    0,    0,    0],
       [   2,  894,   97, ...,    0,    0,    0]])
```

In [104]:

```
import numpy as np
```

In [105]:

```
embeddings_dictionary = dict()
embedding_dim = 100

# Load GloVe 100D embeddings
with open('glove.6B.100d.txt', encoding="utf8") as fp:
    for line in fp.readlines():
        records = line.split()
        word = records[0]
        vector_dimensions = np.asarray(records[1:], dtype='float32')
        embeddings_dictionary [word] = vector_dimensions
```

In [106]:

```
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()

# Train the model
nb.fit(x_train_dtm, y_train)
```

Out[106]:

```
MultinomialNB()
```

In [107]:

```
y_pred_class = nb.predict(x_test_dtm)
y_pred_prob = nb.predict_proba(x_test_dtm)[:, 1]
```

In [108]:

```
from sklearn import metrics
print(metrics.accuracy_score(y_test, y_pred_class))
```

```
0.7141738449490994
```

In [112]:

```
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import Pipeline

pipe = Pipeline([('bow', CountVectorizer()),
                 ('tfid', TfidfTransformer()),
                 ('model', MultinomialNB())])
```

In [113]:

```
pipe.fit(x_train, y_train)

y_pred_class = pipe.predict(x_test)

print(metrics.accuracy_score(y_test, y_pred_class))
```

```
0.7032106499608457
```

In [114]:

```
from sklearn.preprocessing import LabelEncoder
```

In [115]:

```
le = LabelEncoder()
y_encoded = le.fit_transform(y)
```

In [116]:

```
X_train, X_test, y_train, y_test = train_test_split(x, y_encoded, test_size=0.2, random_state=42)
```

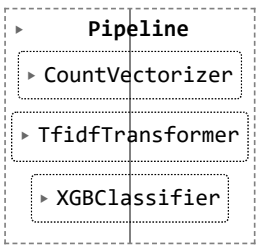
In [117]:

```
import xgboost as xgb

pipe = Pipeline([
    ('bow', CountVectorizer()),
    ('tfid', TfidfTransformer()),
    ('model', xgb.XGBClassifier(
        learning_rate=0.1,
        max_depth=7,
        n_estimators=80,
        use_label_encoder=False,
        eval_metric='auc',
    ))
])
```

In [118]:

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

Out[118]:



In [119]:

```
y_pred = pipe.predict(X_test)
```

In [121]:

```
from sklearn.metrics import accuracy_score
acc = accuracy_score(y_test, y_pred)
print('Test accuracy:', acc)
```

```
Test accuracy: 0.7169441723800196
```