Adipurush Sentiment Analysis: Harnessing Machine Learning to Understand Twitter Buzz

In [3]:
```
import pandas as pd
```

In [4]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [5]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [6]:
```
df = pd.read_csv('adipurush_tweets.csv')
```

In [7]:
```
1 df
```

Out[7]:

	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 [8]:
```
1 df.shape
```

Out[8]:
```
(10001, 4)
```

In [9]:
```
1 df.columns
```

Out[9]:
```
Index(['Date Created', 'Number of Likes', 'Source of Tweet', 'Tweets'], d
type='object')
```

In [10]:
```
1 df.duplicated().sum()
```

Out[10]:
```
1
```

In [11]:
```
1 df = df.drop_duplicates()
```

In [12]:
```
1 df.isnull().sum()
```

Out[12]:
```
Date Created          0
Number of Likes       0
Source of Tweet    10000
Tweets                0
dtype: int64
```

In [13]:
```
1 df = df.drop('Source of Tweet', axis = 1)
```

In [14]:
```
1 df
```

Out[14]:

	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 [15]:
```
1 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 [16]:
```
1 df.describe()
```

Out[16]:

	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 [17]:
```
1 df.nunique()
```

Out[17]:
```
Date Created       9831
Number of Likes     718
Tweets             9874
dtype: int64
```

In [18]:
```
1 df_sorted = df.sort_values(by='Number of Likes', ascending=False)
```

In [19]:
```
1 df_sorted.head(10)
```

Out[19]:

	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 [20]:
```
df['Date Created'] = pd.to_datetime(df['Date Created'])
```

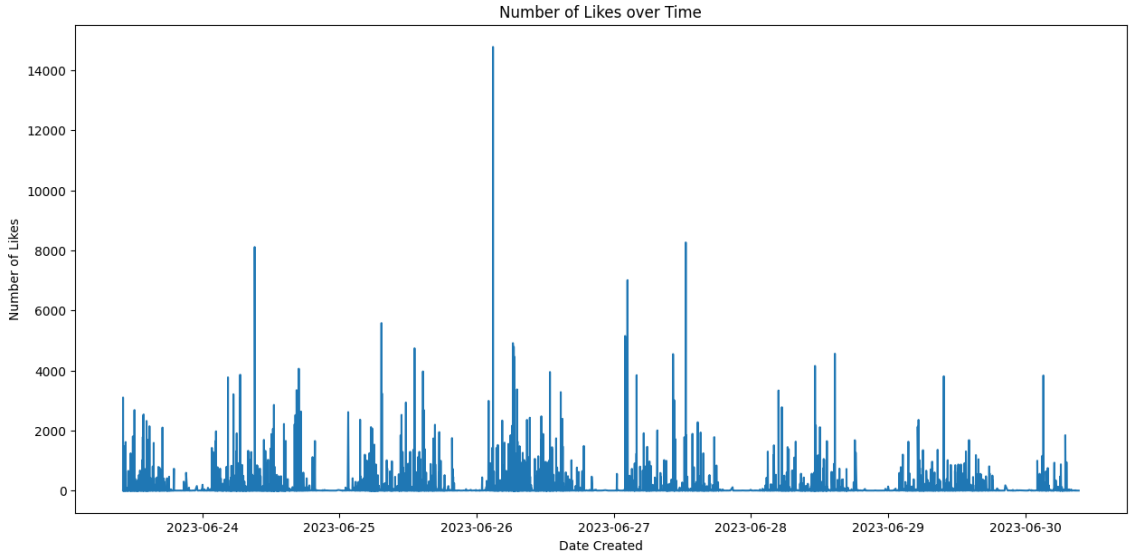
In [21]:
```
df
```

Out[21]:

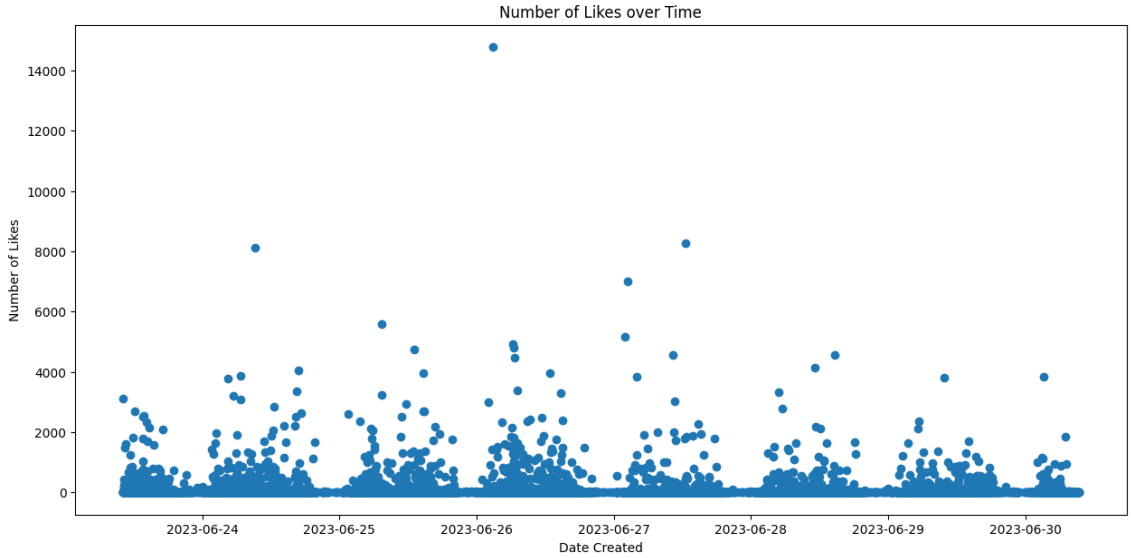
	Date Created	Number of Likes	Tweets
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥🔥🎟 \n@go...
2	2023-06-30 09:20:22+00:00	0	@ponilemova #Adipurush
3	2023-06-30 09:20:00+00:00	3	Adipurush VS 72 Hoorain 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 [22]:
```
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 Lik
plt.xlabel('Date Created')
plt.ylabel('Number of Likes')
plt.title('Number of Likes over Time')
plt.show()
```



In [23]:
```
plt.figure(figsize=[15,7],)
plt.scatter(df_sorted_date['Date Created'], df_sorted_date['Number of
plt.xlabel('Date Created')
plt.ylabel('Number of Likes')
plt.title('Number of Likes over Time')
plt.show()
```



In [24]:
```
import re
import string
from tqdm.notebook import tqdm
from datetime import datetime
import dateutil.parser
```

In [25]:
```
import nltk
from spellchecker import spellchecker
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
```

In [26]:
```
from wordcloud import wordcloud, ImageColorGenerator
from nltk.corpus import stopwords
import random
```

In [27]:
```
nltk.download('vader_lexicon')
nltk.download('stopwords')
```

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

Out[27]:
```
True
```

In [28]:
```
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', 'cata
lan', 'chinese', 'danish', 'dutch', 'english', 'finnish', 'french', 'germ
an', 'greek', 'hebrew', 'hinglish', 'hungarian', 'indonesian', 'italian',
'kazakh', 'nepali', 'norwegian', 'portuguese', 'romanian', 'russian', 'sl
ovene', 'spanish', 'swedish', 'tajik', 'turkish']
```

In [29]:
```
from nltk.tokenize import TweetTokenizer
```

In [30]:
```
english_stopwords = stopwords.words('english')
hinglish_stopwords = stopwords.words('hinglish')
```

In [31]:

```
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, s
    tokens = tokenizer.tokenize(tweet)

    # Remove stopwords for English and Hinglish
    tokens = [token for token in tokens if token not in english_stopwo

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

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

    return cleaned_tweet
```

In [32]:

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

In [33]:
```
df
```

Out[33]:

	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 [34]:
```
def clean_text(text):
    text = text.lower()
    return text.strip()
```

In [35]:
```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: clean_text(x))
```

In [36]:
```
def tokenization(text):
    tokens = re.split('W+',text)
    return tokens
```

In [37]:
```
df.Cleaned_Tweets = df.Cleaned_Tweets.apply(lambda x: tokenization(x))
```

In [38]:
```
from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()
```

In [39]:
```
1 nltk.download('wordnet')
```

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

Out[39]: True

In [40]:
```
1 nltk.download('omw-1.4')
```

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

Out[40]: True

In [41]:
```
1 def lemmatizer (text):
2     lemm_text="".join([wordnet_lemmatizer.lemmatize(word) for word in
3     return lemm_text
```

In [42]:
```
1 df.Cleaned_Tweets=df.Cleaned_Tweets.apply(lambda x: lemmatizer(x))
```

In [43]:
```
1 def remove_digits(text):
2     clean_text = re.sub(r"\b[0-9]+\b\s*", "", text)
3     return(text)
```

In [44]:
```
1 df.Cleaned_Tweets=df.Cleaned_Tweets.apply(lambda x: remove_digits(x))
```

In [45]:
```
1 def remove_digits1(sample_text):
2     clean_text = " ".join([w for w in sample_text.split() if not w.isd
3     return(clean_text)
```

In [46]:
```
1 df.Cleaned_Tweets=df.Cleaned_Tweets.apply(lambda x: remove_digits1(x))
```

In [47]:
```
1  from langdetect import detect
2
3  def detect_language(text):
4      try:
5          lang = detect(text)
6          return lang
7      except:
8          return None
9
10 df['Language'] = df['Cleaned_Tweets'].apply(detect_language)
```

In [48]:
```
df
```

Out[48]:

	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/lsGKcgQuKL	st	no
...	...	...	...	...	...
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...	en
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 [49]:
```
df1 = df.copy()
```

In [50]:
```
df1['english_tweets'] = df[df['Language'] == 'en']['Cleaned_Tweets']
```

In [51]:
```
1 df1
```

Out[51]:

	Date Created	Number of Likes	Tweets	Cleaned_Tweets	Language	englisl
0	2023-06-30 09:21:00+00:00	0	#ENGvAUS #ENGvsAUS #AUSvENG #AUSvsENG #Adipuru...	womens ashes live streaming broadcast tv chann...	en	wome live s bro
1	2023-06-30 09:20:57+00:00	0	Now Playing!! Book Your Ticket Now!! 🎥🏠🎟 \n@go...	playing book ticket	en	play
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	adip h ke co
4	2023-06-30 09:15:22+00:00	3	ST: #Adipurush https://t.co/lsGKcgQuKL	st	no	
...	...	...	...	...	...	
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...	en	adip wee collect
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	em lyric spiri
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	servic bene
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 epic wort

10000 rows × 6 columns

In [52]:
```
1 df1 = df1.dropna()
```

In [53]:
```
1 df1
```

Out[53]:

	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...
...	...	...	...	...	...	...
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...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...	adipurush 1st week ww box office collections ...	en	adipurush 1st week ww box office collections ...
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...

5088 rows × 6 columns

In [54]:
```
df1['Year'] = df1['Date Created'].dt.year
df1['Month'] = df1['Date Created'].dt.month
df1['Day'] = df1['Date Created'].dt.day
```

In [55]:
```
df1
```

Out[55]:

	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...
...	...	...	...	...	...	...
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...
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...	en	adipurush 1st week ww box office collections ఆ...
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...

5088 rows × 9 columns

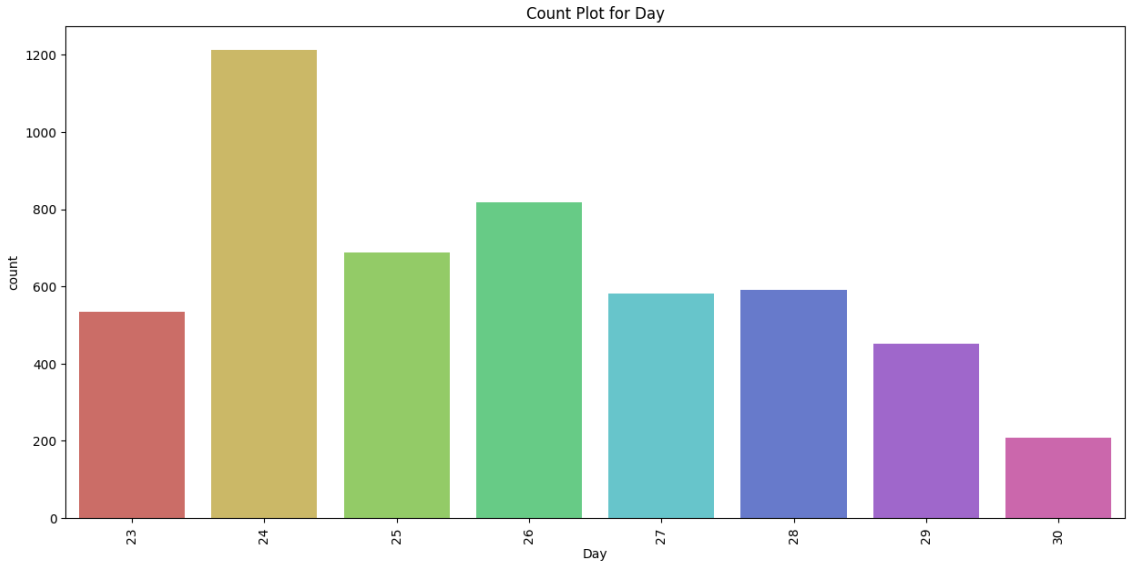
In [56]:
```
1 df1.nunique()
```

Out[56]:
```
Date Created       5051
Number of Likes     555
Tweets             5041
Cleaned_Tweets     4456
Language              1
english_tweets     4456
Year                  1
Month                 1
Day                   8
dtype: int64
```

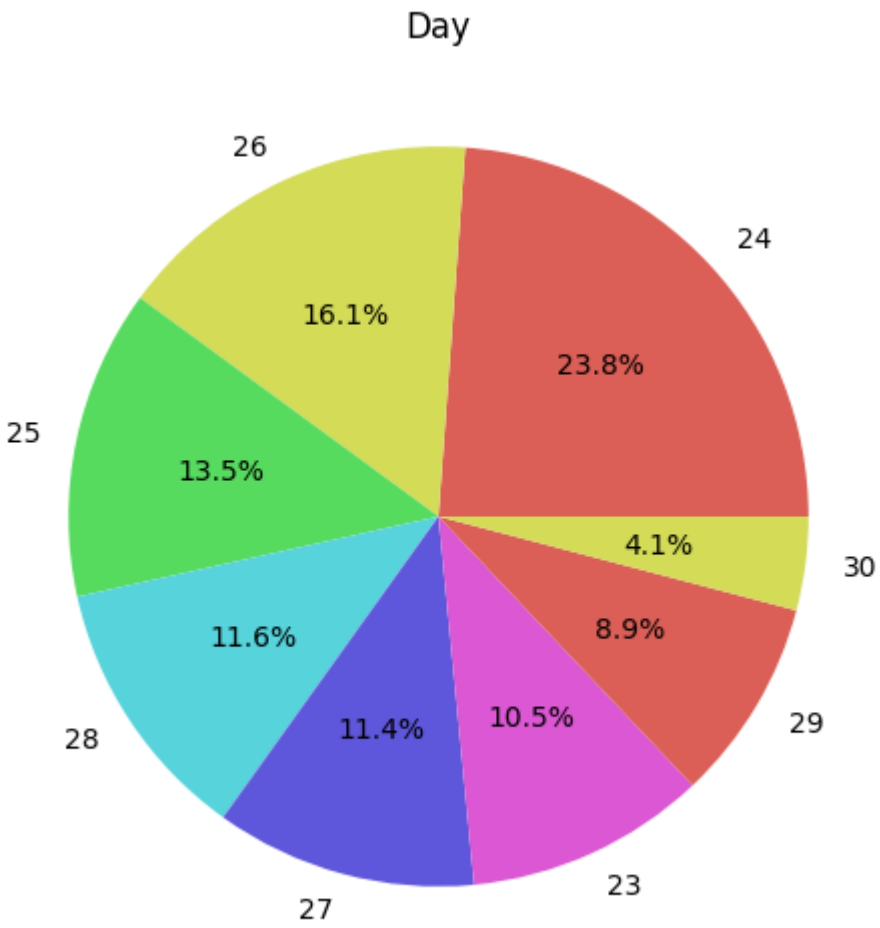
In [57]:
```
1 df1['Time'] = df1['Date Created'].dt.time
```

In [58]:
```
1 df1['Tweet_Length'] = df1['english_tweets'].str.len()
```

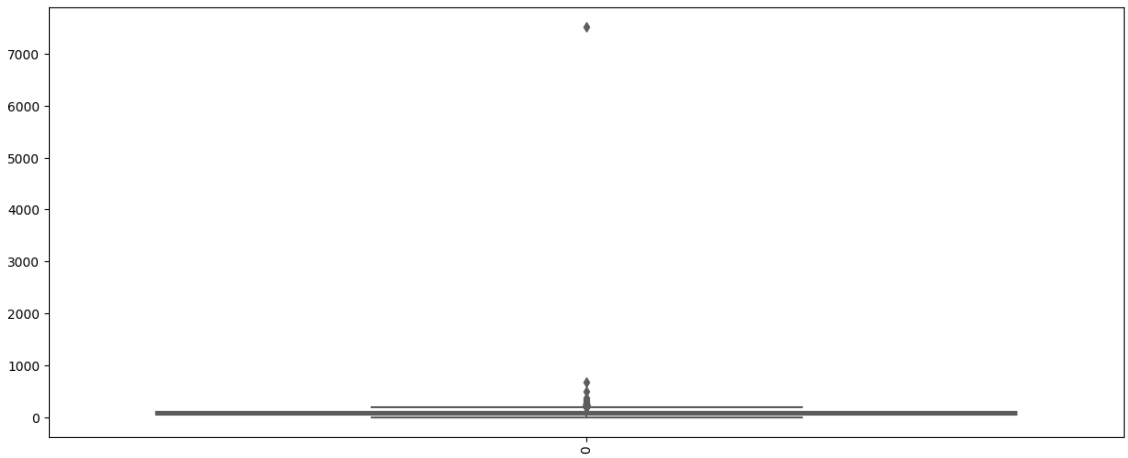
In [59]:
```
1 plt.figure(figsize=[15,7],)
2 plt.title('Count Plot for Day')
3 sns.countplot(x = 'Day', data = df1, palette = 'hls')
4 plt.xticks(rotation = 90)
5 plt.show()
```



In [60]:
```
plt.figure(figsize=(15, 6))
counts = df1['Day'].value_counts()
plt.pie(counts, labels=counts.index, autopct='%1.1f%%', colors=sns.col
plt.title('Day')
plt.show()
```



In [69]:
```
plt.figure(figsize=(15,6))
sns.boxplot(df1['Tweet_Length'],palette='hls')
plt.xticks(rotation=90)
plt.show()
```



In [70]:
```
def label_sentiment(x:float):
    if x < -0.05 : return 'negative'
    if x > 0.35 : return 'positive'
    return 'neutral'
```

In [71]:
```
sia = SIA()
```

In [72]:
```
df1['sentiment'] = [sia.polarity_scores(x)['compound'] for x in tqdm(d
df1['overall_sentiment'] = df1['sentiment'].apply(label_sentiment);
```

100% 5088/5088 [00:01<00:00, 3896.58it/s]

In [74]:
```
1 df1
```

Out[74]:

	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...
...	...	...	...	...	...	...
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...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...	adipurush 1st week ww box office collections ...	en	adipurush 1st week ww box office collections ...
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...

5088 rows × 13 columns

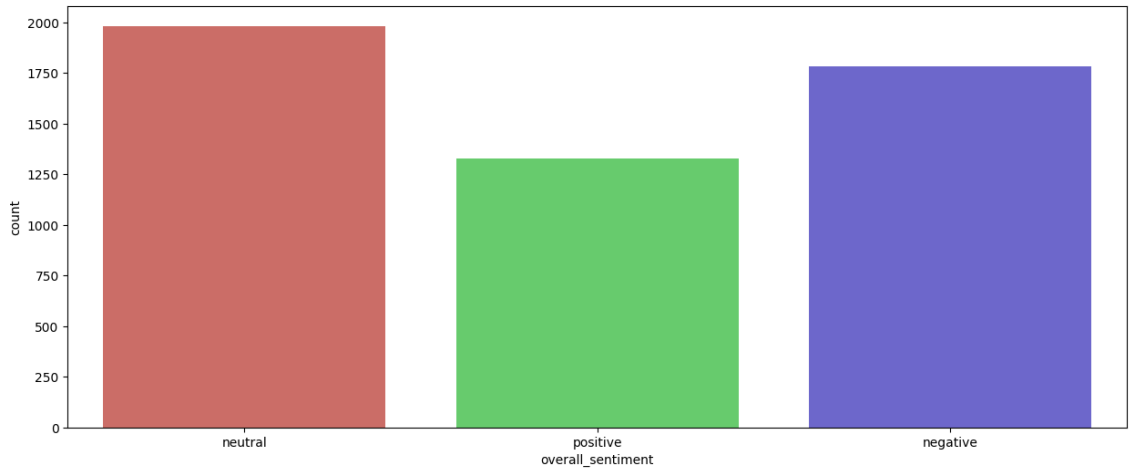
In [75]:
```
1 df1['overall_sentiment'].unique()
```

Out[75]:
```
array(['neutral', 'positive', 'negative'], dtype=object)
```

In [76]:
```
df1['overall_sentiment'].value_counts()
```

Out[76]:
```
neutral     1982
negative    1781
positive    1325
Name: overall_sentiment, dtype: int64
```

In [77]:
```
plt.figure(figsize=(15,6))
sns.countplot(x='overall_sentiment', data = df1, palette = 'hls')
plt.xticks(rotation = 0)
plt.show()
```



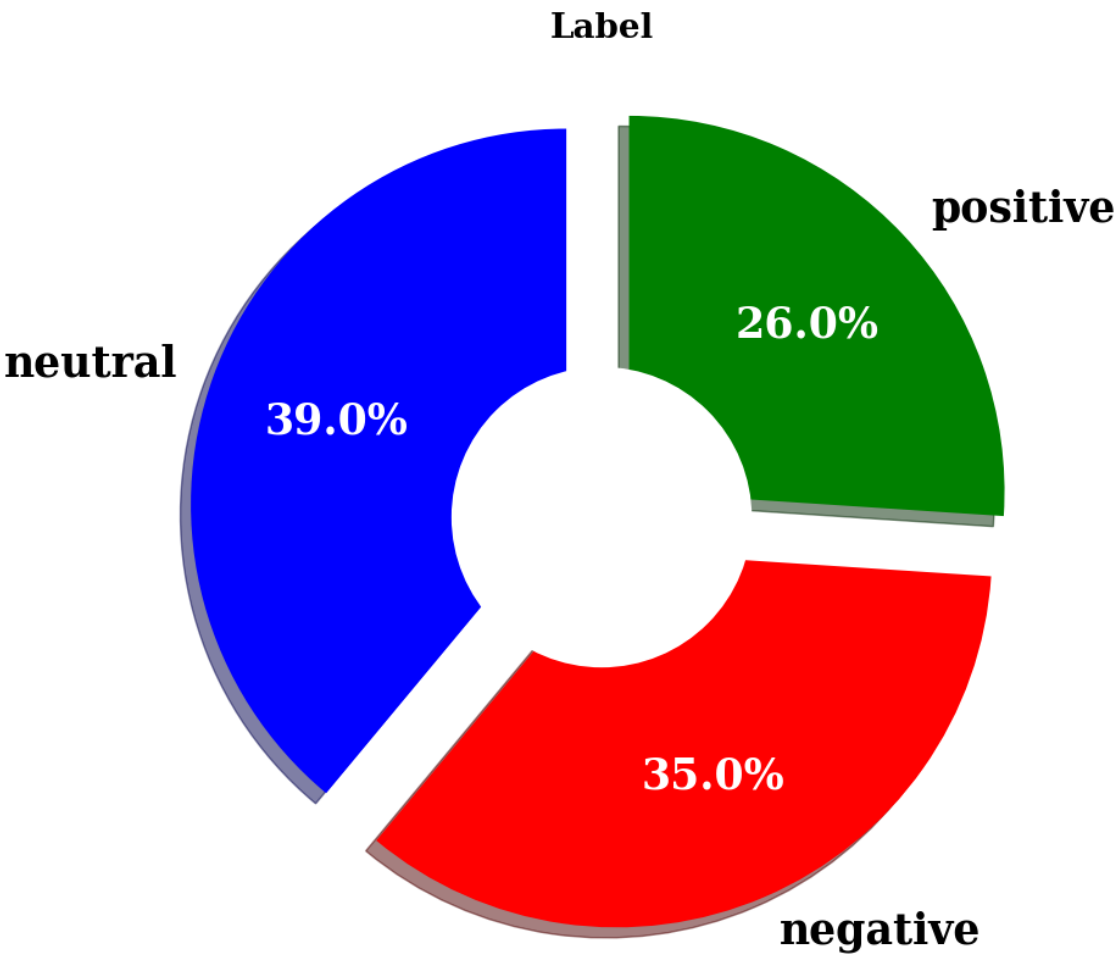
In [78]:

```
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 [79]:
```
1 df1
```

Out[79]:

	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...
...	...	...	...	...	...	...
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...
9996	2023-06-23 10:08:50+00:00	0	Adipurush 1st Week WW Box Office Collections: ...	adipurush 1st week ww box office collections ఌ...	en	adipurush 1st week ww box office collections ఌ...
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...

5088 rows × 13 columns

In [80]:
```
1 df2 = df1[['english_tweets', 'overall_sentiment']]
```

In [81]:
```
df2
```

Out[81]:

	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
...	...	...
9995	rangarajan garu poojari chilkur balaji appreci...	positive
9996	adipurush 1st week ww box office collections	neutral
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

5088 rows × 2 columns

In [82]:
```
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 [83]:
```
1 df2
```

Out[83]:

	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...
...	...	...	...
9995	rangarajan garu poojari chilkur balaji appreci...	positive	rangarajan garu poojari chilkur balaji appreci...
9996	adipurush 1st week ww box office collections ...	neutral	adipurush st week ww box office collections st
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...

5088 rows × 3 columns

In [84]:
```
1 df3 = df2[['Cleaned_English_Tweets', 'overall_sentiment']]
```

In [85]:
```
df3
```

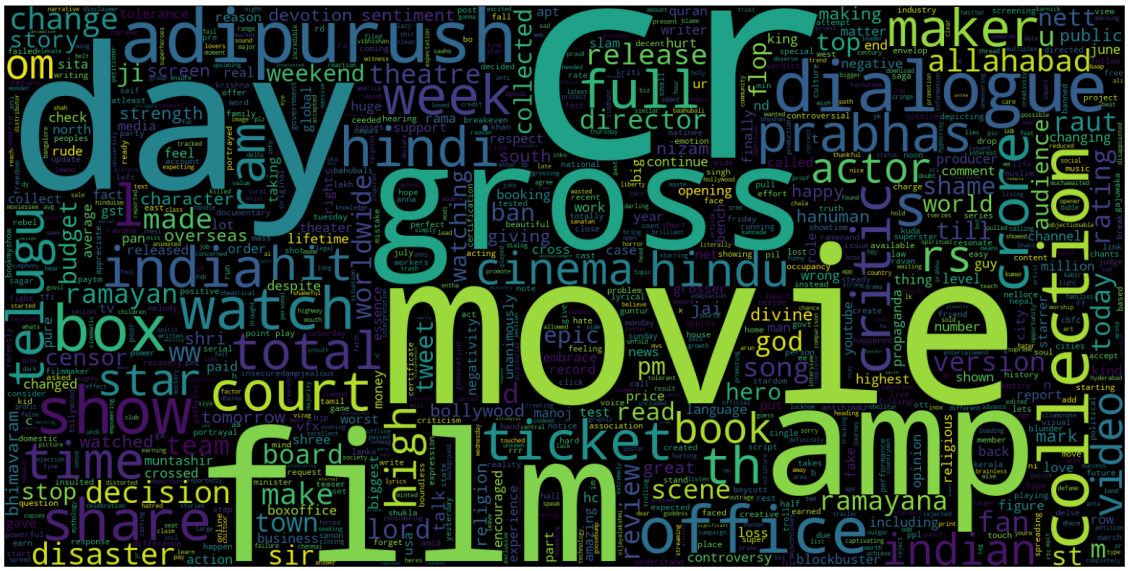
Out[85]:

	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
...	...	...
9995	rangarajan garu poojari chilkur balaji appreci...	positive
9996	adipurush st week ww box office collections st	neutral
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

5088 rows × 2 columns

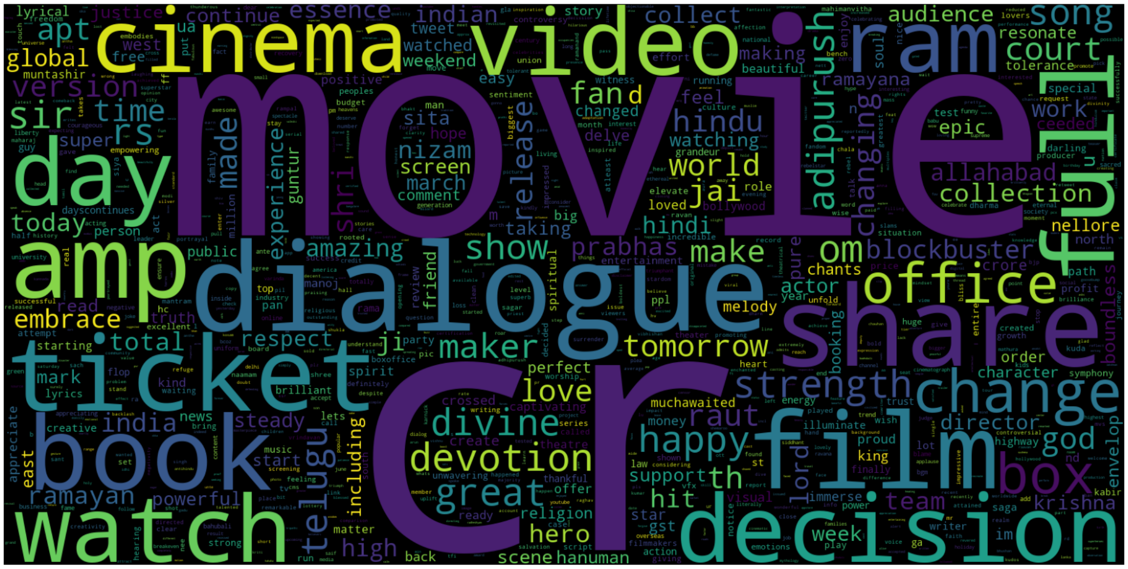
In [86]:
```
import wordcloud
```

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]:

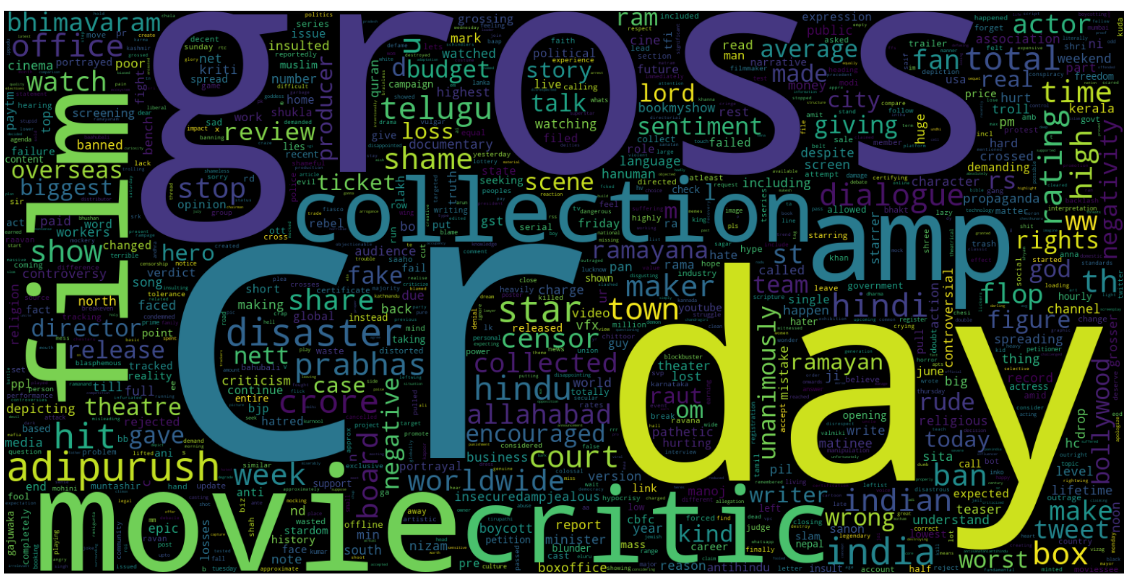
```
data = df3[df3['overall_sentiment']=="positive"]['Cleaned_English_Twee
plt.figure(figsize = (20,20))
wc = WordCloud(max_words = 1000 , width = 1600 , height = 800,
               collocations=False).generate(" ".join(data))
plt.imshow(wc)
plt.axis('off')
plt.show()
```



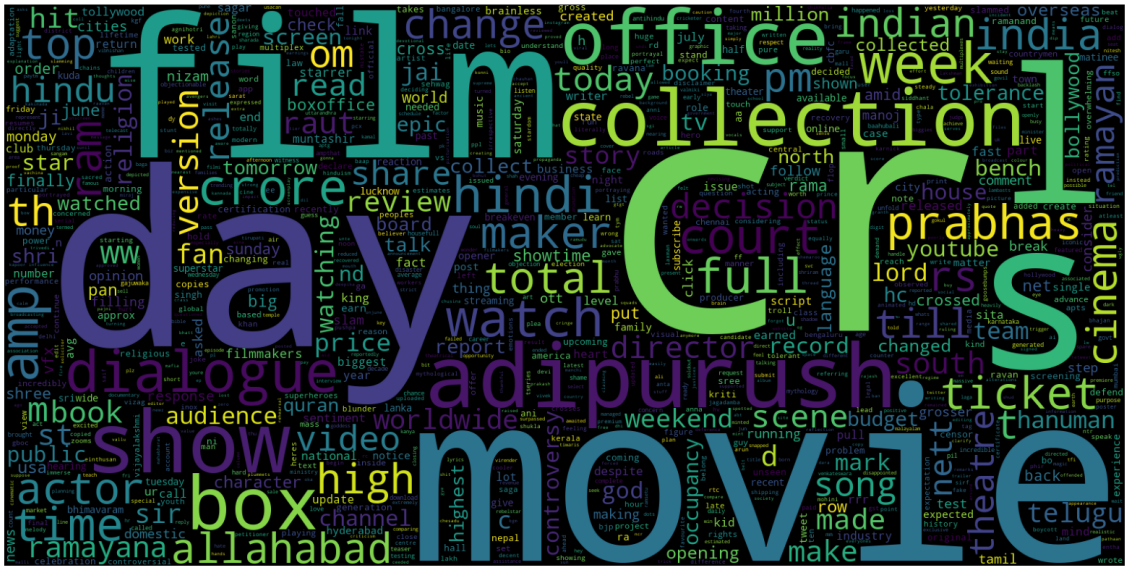
In [89]:

```

data = df3[df3['overall_sentiment']=="negative"]['Cleaned_English_Twee
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 [90]:
```
data = df3[df3['overall_sentiment']=="neutral"]['Cleaned_English_Tweet
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 [91]:
```
# Load the libraries required for performing classification

from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix, accuracy_score, f1_score
from sklearn.model_selection import train_test_split
from nltk.tokenize import TweetTokenizer
from sklearn.feature_extraction.text import CountVectorizer, TfidfTran
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [92]:
```
# Split the data into training and testing data sets
# Use cleaned english as independent variable and overall sentiment as

X = df3["Cleaned_English_Tweets"].values
y = df3['overall_sentiment'].values

X_train, X_test, y_train, y_test = train_test_split(X,y, random_state=
```

In [93]:
```
# Extract features using TFIDF Vectorizer

vectorizer = TfidfVectorizer(max_features=1000)
X_train_idf = vectorizer.fit_transform(X_train)
X_test_idf = vectorizer.transform(X_test)
```

In [94]:
```
# Print idf values
df_idf = pd.DataFrame(vectorizer.idf_, index=vectorizer.get_feature_na
# Sort ascending
df_idf.sort_values(by=['idf_weights'],ascending = False).head()
```

Out[94]:

	idf_weights
narratives	7.791783
kashmir	7.791783
artistic	7.791783
sree	7.791783
lks	7.568640

In [95]:
```
# Perform Multinomial Naive Bayes Classification
# Apply MultinomialNB on training data
mnb = MultinomialNB()
mnb.fit(X_train_idf, y_train)
```

Out[95]:
```
MultinomialNB
MultinomialNB()
```

In [96]:
```
# Predict polarity by fitting the model to testing data
pred_mnb = mnb.predict(X_test_idf)

# Calculate accuracy of predicted values
acc = accuracy_score(y_test, pred_mnb)


results = pd.DataFrame([['Multinomial Naive Bayes', acc]],
               columns = ['Model', 'Accuracy'])

print(results)
```

```
                   Model  Accuracy
0  Multinomial Naive Bayes  0.667322
```

In [97]:
```
# Perform Random Forest classification on the processed data and compa

# Random Forest Classifier with 'gini'

from sklearn.ensemble import RandomForestClassifier
clf_rf = RandomForestClassifier()
clf_rf.fit(X_train_idf, y_train)

# Predict using testing data
y_pred_rf = clf_rf.predict(X_test_idf)

# Calculate accuracy
acc = accuracy_score(y_test, y_pred_rf)

model_results = pd.DataFrame([['Random Forest(Gini)', acc]],
               columns = ['Model', 'Accuracy'])

results = results.append(model_results, ignore_index = True)
print(results)
```

```
                   Model  Accuracy
0  Multinomial Naive Bayes  0.667322
1      Random Forest(Gini)  0.747217
```

In [98]:
```
# Random Forest Classifier with 'entropy'

from sklearn.ensemble import RandomForestClassifier
clf_rf = RandomForestClassifier(criterion='entropy')
clf_rf.fit(X_train_idf, y_train)

# Predict using testing data
y_pred_rf = clf_rf.predict(X_test_idf)

# Calculate accuracy
acc = accuracy_score(y_test, y_pred_rf)

model_results = pd.DataFrame([['Random Forest(Entropy)', acc]],
               columns = ['Model', 'Accuracy'])

results = results.append(model_results, ignore_index = True)
print(results)
```

```
                   Model  Accuracy
0  Multinomial Naive Bayes  0.667322
1      Random Forest(Gini)  0.747217
2   Random Forest(Entropy)  0.747872
```

In [99]:
```
# Display confusion matrix for Random Forest

confusion_matrix(y_test,y_pred_rf) ### Confusion matrix for Random For
```

Out[99]:
```
array([[386, 124,   7],
       [ 46, 510,  46],
       [ 32, 130, 246]], dtype=int64)
```

In [ ]:
```
1
```

In [ ]:
```
1
```