```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import nltk
from sklearn.feature_extraction.text import CountVectorizer
from wordcloud import WordCloud,STOPWORDS
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
import re,string,unicodedata
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score,f1_score
from sklearn.model_selection import train_test_split
from string import punctuation
from nltk import pos_tag
from nltk.corpus import wordnet
import re
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
```

```
df=pd.read_csv('/content/drive/MyDrive/Sentiment Analysis/Data/IMDB-Dataset.csv', encoding='latin-1')
```

```
#Customize stopword as per data
import nltk
nltk.download('stopwords')
stop_words = stopwords.words('english')
new_stopwords = ["would","shall","could","might"]
stop_words.extend(new_stopwords)
stop_words.remove("not")
stop_words=set(stop_words)
print(stop_words)
```

```
{'weren', 'we', 'below', 'there', 'needn', 'the', 'that', 'an', 'each', 'is', 'those', 'himself', 'yourselves', 'again', 'all', 'be
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
#Removing special character
def remove_special_character(content):
    return re.sub('\W+',' ', content )#re.sub('\[[^&@#!]]*\]', '', content)

# Removing URL's
def remove_url(content):
    return re.sub(r'http\S+', '', content)

#Removing the stopwords from text
def remove_stopwords(content):
    clean_data = []
    for i in content.split():
        if i.strip().lower() not in stop_words and i.strip().lower().isalpha():
            clean_data.append(i.strip().lower())
    return " ".join(clean_data)

# Expansion of english contractions
def contraction_expansion(content):
    content = re.sub(r"won\'t", "would not", content)
    content = re.sub(r"can\'t", "can not", content)
    content = re.sub(r"don\'t", "do not", content)
    content = re.sub(r"shouldn\'t", "should not", content)
    content = re.sub(r"needn\'t", "need not", content)
    content = re.sub(r"hasn\'t", "has not", content)
    content = re.sub(r"haven\'t", "have not", content)
    content = re.sub(r"weren\'t", "were not", content)
    content = re.sub(r"mightn\'t", "might not", content)
    content = re.sub(r"didn\'t", "did not", content)
```

```
    content = re.sub(r"n\'t", " not", content)
    '''content = re.sub(r"\'re", " are", content)
    content = re.sub(r"\'s", " is", content)
    content = re.sub(r"\'d", " would", content)
    content = re.sub(r"\'ll", " will", content)
    content = re.sub(r"\'t", " not", content)
    content = re.sub(r"\'ve", " have", content)
    content = re.sub(r"\'m", " am", content)'''
    return content

#Data preprocessing
def data_cleaning(content):
    content = contraction_expansion(content)
    content = remove_special_character(content)
    content = remove_url(content)

    content = remove_stopwords(content)
    return content
```

```
pd.options.display.max_colwidth = 1000
#Data cleaning
df['Reviews_clean']=df['Reviews'].apply(data_cleaning)
df.head(5)
```

1	1	as Disaster films? Selzterwater and Failburg once again have shown that they lack any sort of writing skill and humor. Having unfortunately been tortured with Date Movie and Epic Movie I know exactly what to expect from these two...no plot, no jokes just bad references and cheaply remade scenes from other films. Someone should have informed them that satire is more than just copy and paste from one film to another, though I shouldn't say that because some of these actually just seem to be taken from trailers.There is nothing clever or witty or re...	Disaster Movie	Winehouse ou Hancock se registram como filmes de Desastre? Selzterwater e Failburg mostraram mais uma vez que nÃ£o possuem nenhum tipo de habilidade e humor de escrita. Infelizmente, tendo sido torturado com Date Movie e Epic Movie, sei exatamente o que esperar desses dois ... nenhum enredo, nenhuma piada, apenas mÃ¡s referÃªncias e cenas refeitas de outros filmes. AlguÃ©m deveria ter informado a eles que a sÃ¡tira Ã© mais do que apenas copiar e colar de um filme para outro, embora eu nÃ£o deva dizer isso porque algu...	two plot jokes bad references cheaply remade scenes films someone informed satire copy paste one film another though not say actually seem taken trailers nothing clever witty remotely smart way two write not believe people still pay see travesties insult audience though enjoy films doubt smart enough realize rating unfortunately not number low enough yes includes negatives rate deserves top worst films time right date movie epic faliure mean movie meet spartans rather forced hour manos hands fate marathon watch slop
		Really, I could write a scathing review of this turd sandwich, but instead, I'm just going to be making a few observations and points I've deduced.There's just no point in watching these movies anymore. Does any reader out there remember Scary Movie? Remember how it was original with a few comedic elements to it? There was slapstick, some funny lines, it was a pretty forgettable comedy, but it was worth the price of		Realmente, eu poderia escrever uma crÃtica contundente sobre esse sanduÃche de cocÃ´, mas, em vez disso, vou fazer algumas observaÃ§Ãµes e pontos que deduzi. NÃ£o hÃ¡ mais sentido assistir a esses filmes. Algum leitor por aÃ se lembra do filme de terror? Lembra como era original, com alguns elementos cÃ´micos? Havia palhaÃ§ada, algumas frases engraÃ§adas, era uma comÃ©dia bastante esquecÃvel, mas valia o preÃ§o da entrada. Bem, essa	really write scathing review turd sandwich instead going making observations points deduced point watching movies anymore reader remember scary movie remember original comedic elements slapstick funny lines pretty forgettable comedy worth price admission well last

```
#Mapping rating data to Binary label 1 (+ve) if rating >=7 and 0 (-ve) if rating <=4 and 2 (neutral) if ra
df['Label'] = df['Ratings'].apply(lambda x: '1' if x >= 7 else ('0' if x<=4 else '2'))
#Removing
df=df[df.Label<'2']
data=df[['Reviews_clean','Reviews','Ratings','Label']]
print(data['Label'].value_counts())
data.tail(5)
```

```
Name: Label, dtype: int64
```

	Reviews_clean	Reviews	Ratings	Label
149995	goldeneye number personal favorite movie time james bond action film best one ever made opinion best one love movie death grew watching child pierce brosnan number favorite james bond goldeneye first film star pierce brosnan fictional officer james bond film directed martin campbell first series not take story elements works novelist ian fleming hard core james bond fan make apologies believing pierce brosnan closest thing seen ian fleming james bond reason favorite feel action action action action never feel bored long like bond movies fast paced entertaining hard core action plot simply story never get bored beautiful movie beautiful direction martin campbell stunts	GoldenEye (1995) is my number 1 personal favorite movie of all time! James Bond 007 action film the best one that was ever made. In my opinion it is the best one I love this movie to death! I grew up watching it as a child and Pierce Brosnan is my number 1 favorite James Bond. GoldenEye is the first film to star Pierce Brosnan as the fictional MI6 officer James Bond. The film was directed by Martin Campbell and is the first in the series not to take story elements from the works of novelist Ian Fleming.I'm a hard-core James Bond fan. I make no apologies for believing that Pierce Brosnan is the closest thing we've seen to IAN FLEMING's	10	1

Character level tokenization

goldeneye performed tina turner goldeneye

```
def characterTokenization(content):
  character_tokenized_data=[]
  special_removed=remove_special_character(content)
  for character in special_removed:
    character_tokenized_data.append(character)

  return character_tokenized_data
```

149996	without fans needed movie like everything	without 007 the fans needed a movie like	10	1

```
df['Reviews_Character_Tokenized']=df['Reviews'].apply(characterTokenization)
df.head(5)
```

		remember Scary Movie? Remember how it was original with a few comedic elements to it? There was slapstick, some funny lines, it was a pretty forgettable comedy, but it was worth the price of admission. Well, That was the last time this premise was funny. STOP MAKING THESE MOVIES. PLEASE.I could call for a boycott of these pieces of monkey sh*t, but we all know there's going to be a line up of pre pubescent annoying little buggers, spouting crappy one liners like, "THIS IS SPARTA!" and, "IM RICK JAMES BITCH" so these movies will continue to make some form of monetary gain, considering the production value of this movie looks like it cost about 10 cents to make.Don't see this movie. Don't spend any money on it. Go home, rent Airplane, laugh your ass off, and ...	Disaster Movie	original, com alguns elementos cÃ´micos? Havia palhaÃ§ada, algumas frases engraÃ§adas, era uma comÃ©dia bastante esquecÃ-vel, mas valia o preÃ§o da entrada. Bem, essa foi a Ãºltima vez que essa premissa foi engraÃ§ada. PARE DE FAZER ESTES FILMES. POR FAVOR, eu poderia pedir um boicote a esses pedaÃ§os de macaco, mas todos sabemos que haverÃ¡ uma fila de buggers irritantes e prÃ©-pubescentes, jorrando uns forros ruins como: "ISTO Ã SPARTA!" e "IM RICK JAMES BITCH", para que esses filmes continuem gerando algum ganho monetÃ¡rio, considerando que o valor de produÃ§Ã£o deste filme parece custar cerca de 10 centavos de dÃ³lar. NÃ£o gaste dinheiro com isso. VÃ¡ para casa, alugue a Airplane, ria e julgue silenciosament...	scary movie remember original comedic elements slapstick funny lines pretty forgettable comedy worth price admission well last time premise funny stop making movies please call boycott pieces monkey sh know going line pre pubescent annoying little buggers spouting crappy one liners like sparta im rick james bitch movies continue make form monetary gain considering production value movie looks like cost cents make not see movie not spend money go home rent airplane laugh ass silently judge people talking movie monday favor	
2	1	If you saw the other previous spoof movies by these two horrible gentlemen, then you should know that this already will be bad. I'll tell you the truth, if you want to watch it as a brainless person (ironically meant for the		Se vocÃª viu os outros filmes falsificados anteriores por esses dois senhores horrÃ-veis, deve saber que isso jÃ¡ serÃ¡ ruim. Vou lhe dizer a verdade, se vocÃª quiser vÃª-lo como uma pessoa sem cÃ©rebro (ironicamente para os adolescentes estereotipados, o que eu nÃ£o sou), entÃ£o vocÃª rirÃ¡	saw previous spoof movies two horrible gentlemen know already bad tell truth want watch brainless person ironically meant stereotypical teenagers not laugh bit judge even little movie automatically fails never ask comes two men remember good old hollywood days whenever making movie	[R, e, a, l, l, y, , l, , c, o, u, w, r, i, t, e, , a, , s, c, a, t, g, , r, e, v, i, e, w, , o, f, , t, t, u, r, d, , s, a, n, d, w, i, c, u, t, , i, n, s, t, e, a, d, , l, u, s, t, , g, o, i, n, g, , t, o, m, a, k, i,

Word level Tokenization

```
def wordLevelTokenization(content):
  word_tokenized_data=[]
  word_tokenized_data=content.split()
  return word_tokenized_data
```

```
df['Reviews_Word_Tokenized']=df['Reviews'].apply(wordLevelTokenization)
df.head(5)
```