In [3]:
```
#importing libraries required for the analysis
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
plt.style.use("ggplot")
import re
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
stopwords= set(stopwords.words('english'))
from wordcloud import WordCloud
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
```

In [4]:
```
#Reading the data and finding the first 5 rows of data
df= pd.read_csv('IMDB Dataset.csv')
df.head()
```

Out[4]:

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

In [5]:
```
#Understanding the shape of the data
df.shape
```

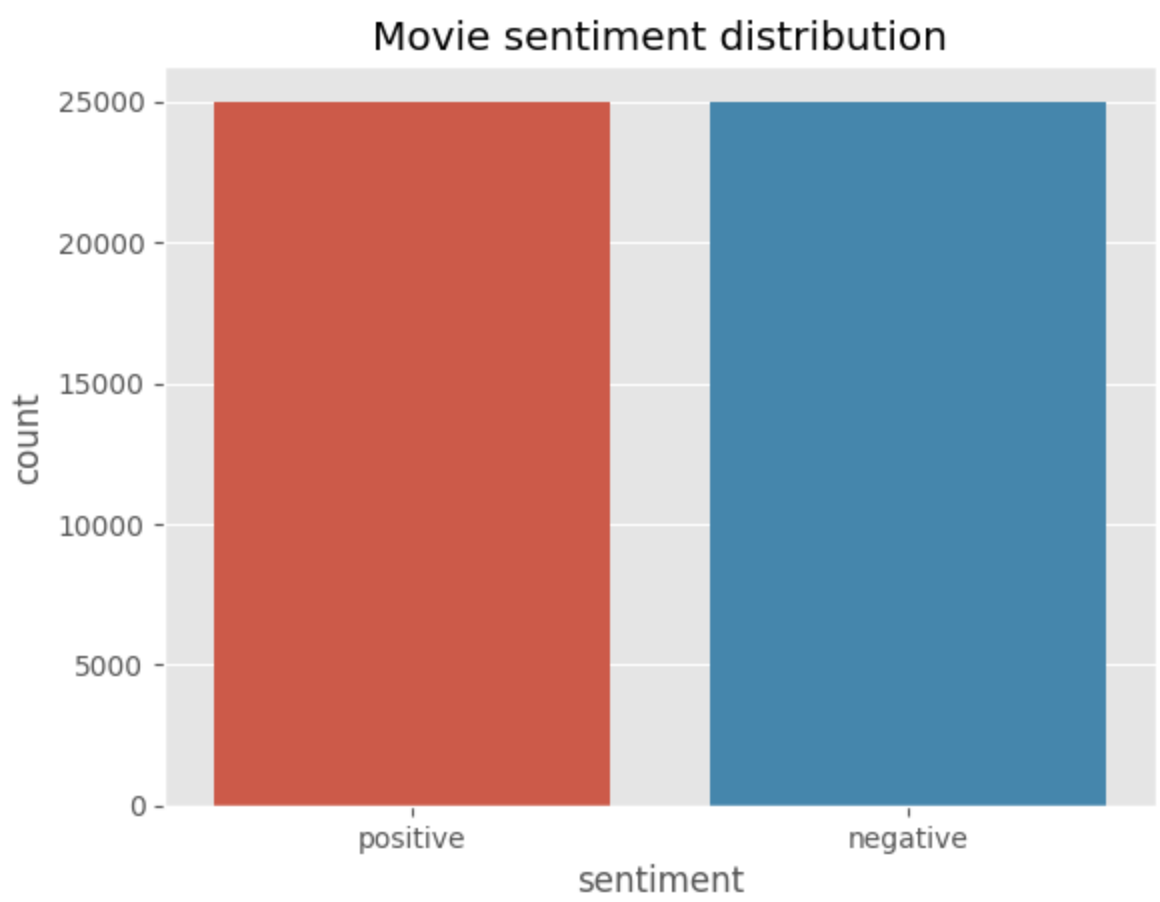
Out[5]:
```
(50000, 2)
```

In [6]:
```
# finding the shape of the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   review     50000 non-null  object
 1   sentiment  50000 non-null  object
dtypes: object(2)
memory usage: 781.4+ KB
```

In [7]:
```
#Plotting the distribution of sentiments in the dataset
sns.countplot(x='sentiment',data=df)
plt.title("Movie sentiment distribution")
```

Out[7]:
```
Text(0.5, 1.0, 'Movie sentiment distribution')
```



In [8]:
```
#Displaying the first five rows
for i in range(5):
    print("Review: ", [i])
    print(df['review'].iloc[i],"\n")
    print("Sentiment: ", df["sentiment"].iloc[i],"\n\n")
```

```
Review:  [0]
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be
hooked. They are right, as this is exactly what happened with me.<br /><br />The first t
hing that struck me about Oz was its brutality and unflinching scenes of violence, which
set in right from the word GO. Trust me, this is not a show for the faint hearted or tim
id. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore,
in the classic use of the word.<br /><br />It is called OZ as that is the nickname given
to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an
experimental section of the prison where all the cells have glass fronts and face inward
s, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangs
tas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy d
ealings and shady agreements are never far away.<br /><br />I would say the main appeal
of the show is due to the fact that it goes where other shows wouldn't dare. Forget pret
ty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't
mess around. The first episode I ever saw struck me as so nasty it was surreal, I could
n't say I was ready for it, but as I watched more, I developed a taste for Oz, and got a
ccustomed to the high levels of graphic violence. Not just violence, but injustice (croo
ked guards who'll be sold out for a nickel, inmates who'll kill on order and get away wi
th it, well mannered, middle class inmates being turned into prison bitches due to their
lack of street skills or prison experience) Watching Oz, you may become comfortable with
what is uncomfortable viewing....thats if you can get in touch with your darker side.

Sentiment:  positive


Review:  [1]
A wonderful little production. <br /><br />The filming technique is very unassuming- ver
y old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of rea
lism to the entire piece. <br /><br />The actors are extremely well chosen- Michael Shee
n not only "has got all the polari" but he has all the voices down pat too! You can trul
```

y see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrificly written and performed piece. A masterful production about one of the great master's of comedy and his life.

The realism really comes home with the little things: the fantasy of the guard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well done.

Sentiment: positive

Review: [2]
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a light-hearted comedy. The plot is simplistic, but the dialogue is witty and the characters are likable (even the well bread suspected serial killer). While some may be disappointed when they realize this is not Match Point 2: Risk Addiction, I thought it was proof that Woody Allen is still fully in control of the style many of us have grown to love.

This was the most I'd laughed at one of Woody's comedies in years (dare I say a decade?). While I've never been impressed with Scarlet Johanson, in this she managed to tone down her "sexy" image and jumped right into a average, but spirited young woman.

This may not be the crown jewel of his career, but it was wittier than "Devil Wears Prada" and more interesting than "Superman" a great comedy to go see with friends.

Sentiment: positive

Review: [3]
Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the time.

This movie is slower than a soap opera... and suddenly, Jake decides to become Rambo and kill the zombie.

OK, first of all when you're going to make a film you must Decide if its a thriller or a drama! As a drama the movie is watchable. Parents are divorcing & arguing like in real life. And then we have Jake with his closet which totally ruins all the film! I expected to see a BOOGEYMAN similar movie, and instead i watched a drama with some meaningless thriller spots.

3 out of 10 just for the well playing parents & descent dialogs. As for the shots with Jake: just ignore them.

Sentiment: negative

Review: [4]
Petter Mattei's "Love in the Time of Money" is a visually stunning film to watch. Mr. Mattei offers us a vivid portrait about human relations. This is a movie that seems to be telling us what money, power and success do to people in the different situations we encounter.

This being a variation on the Arthur Schnitzler's play about the same theme, the director transfers the action to the present time New York where all these different characters meet and connect. Each one is connected in one way, or another to the next person, but no one seems to know the previous point of contact. Stylishly, the film has a sophisticated luxurious look. We are taken to see how these people live and the world they live in their own habitat.

The only thing one gets out of all these souls in the picture is the different stages of loneliness each one inhabits. A big city is not exactly the best place in which human relations find sincere fulfillment, as one discerns is the case with most of the people we encounter.

The acting is good under Mr. Mattei's direction. Steve Buscemi, Rosario Dawson, Carol Kane, Michael Imperioli, Adrian Grenier, and the rest of the talented cast, make these characters come alive.

We wish Mr. Mattei good luck and await anxiously for his next work.

Sentiment: positive

In [9]:
```
#Defining a function to get the word count
def word_count(text):
```

```
    text= text.split()
    word_count= len(text)
    return word_count
```

In [10]:
```
#Finding the word count of the reviews
df["word_count"]= df['review'].apply(word_count)
df.head()
```

Out[10]:

	review	sentiment	word_count
0	One of the other reviewers has mentioned that ...	positive	307
1	A wonderful little production. The...	positive	162
2	I thought this was a wonderful way to spend ti...	positive	166
3	Basically there's a family where a little boy ...	negative	138
4	Petter Mattei's "Love in the Time of Money" is...	positive	230

In [11]:
```
#Printing the review with the maximum word count
pd.set_option('display.max_colwidth', None) #setting the maximum column width to see the
print(df[df["word_count"]==max(df["word_count"])])
```

review \

31481 Match 1: Tag Team Table Match Bubba Ray and Spike Dudley vs Eddie Guerrero and Ch
ris Benoit Bubba Ray and Spike Dudley started things off with a Tag Team Table Match aga
inst Eddie Guerrero and Chris Benoit. According to the rules of the match, both opponent
s have to go through tables in order to get the win. Benoit and Guerrero heated up early
on by taking turns hammering first Spike and then Bubba Ray. A German suplex by Benoit t
o Bubba took the wind out of the Dudley brother. Spike tried to help his brother, but th
e referee restrained him while Benoit and Guerrero ganged up on him in the corner. With
Benoit stomping away on Bubba, Guerrero set up a table outside. Spike dashed into the ri
ng and somersaulted over the top rope onto Guerrero on the outside! After recovering and
taking care of Spike, Guerrero slipped a table into the ring and helped the Wolverine se
t it up. The tandem then set up for a double superplex from the middle rope which would
have put Bubba through the table, but Spike knocked the table over right before his brot
her came crashing down! Guerrero and Benoit propped another table in the corner and trie
d to Irish Whip Spike through it, but Bubba dashed in and blocked his brother. Bubba cau
ght fire and lifted both opponents into back body drops! Bubba slammed Guerrero and Spik
e stomped on the Wolverine from off the top rope. Bubba held Benoit at bay for Spike to
soar into the Wassup! headbutt! Shortly after, Benoit latched Spike in the Crossface, bu

t the match continued even after Spike tapped out. Bubba came to his brother's rescue an d managed to sprawl Benoit on a table. Bubba leapt from the middle rope, but Benoit move d and sent Bubba crashing through the wood! But because his opponents didn't force him t hrough the table, Bubba was allowed to stay in the match. The first man was eliminated s hortly after, though, as Spike put Eddie through a table with a Dudley Dawg from the rin g apron to the outside! Benoit put Spike through a table moments later to even the scor e. Within seconds, Bubba nailed a Bubba Bomb that put Benoit through a table and gave th e Dudleys the win! Winner: Bubba Ray and Spike Dudley

Match 2: Cruiserweight Championship Jamie Noble vs Billy Kidman Billy Kidman challenged Jamie Noble, who brough t Nidia with him to the ring, for the Cruiserweight Championship. Noble and Kidman locke d up and tumbled over the ring, but raced back inside and grappled some more. When Kidma n thwarted all Noble's moves, Noble fled outside the ring where Nidia gave him some enco uragement. The fight spread outside the ring and Noble threw his girlfriend into the cha llenger. Kidman tossed Nidia aside but was taken down with a modified arm bar. Noble con tinued to attack Kidman's injured arm back in the ring. Kidman's injured harm hampered h is offense, but he continued to battle hard. Noble tried to put Kidman away with a power bomb but the challenger countered into a facebuster. Kidman went to finish things with a Shooting Star Press, but Noble broke up the attempt. Kidman went for the Shooting Star P ress again, but this time Noble just rolled out of harm's way. Noble flipped Kidman into a power bomb soon after and got the pin to retain his WWE Cruiserweight Championship! Wi nner: Jamie Noble

Match 3: European Championship William Regal vs Jeff Hardy William Regal took on Jeff Hardy next in an attempt to win back the European Championshi p. Jeff catapulted Regal over the top rope then took him down with a hurracanrana off th e ring apron. Back in the ring, Jeff hit the Whisper in the wind to knock Regal for a lo op. Jeff went for the Swanton Bomb, but Regal got his knees up to hit Jeff with a devast ating shot. Jeff managed to surprise Regal with a quick rollup though and got the pin to keep the European Championship! Regal started bawling at seeing Hardy celebrate on his w ay back up the ramp. Winner: Jeff Hardy

Match 4: Chris Jericho vs John Cena C hris Jericho had promised to end John Cena's career in their match at Vengeance, which c ame up next. Jericho tried to teach Cena a lesson as their match began by suplexing him to the mat. Jericho continued to knock Cena around the ring until his cockiness got the better of him. While on the top rope, Jericho began to showboat and allowed Cena to grab him for a superplex! Cena followed with a tilt-a-whirl slam but was taken down with a na sty dropkick to the gut. The rookie recovered and hit a belly to belly suplex but could n't put Y2J away. Jericho launched into the Lionsault but Cena dodged the move. Jericho nailed a bulldog and then connected on the Lionsault, but did not go for the cover. He g oaded Cena to his feet so he could put on the Walls of Jericho. Cena had other ideas, re versing the move into a pin attempt and getting the 1-2-3! Jericho went berserk after th e match. Winner: John Cena

Match 5: Intercontinental Championship RVD vs Broc k Lesnar via disqualification The Next Big Thing and Mr. Pay-Per-View tangled with the I ntercontinental Championship on the line. Brock grabbed the title from the ref and drape d it over his shoulder momentarily while glaring at RVD. Van Dam 's quickness gave Brock fits early on. The big man rolled out of the ring and kicked the steel steps out of frus tration. Brock pulled himself together and began to take charge. With Paul Heyman beamin g at ringside, Brock slammed RVD to the hard floor outside the ring. From there, Brock b egan to overpower RVD, throwing him with ease over the top rope. RVD landed painfully on his back, then had to suffer from having his spine cracked against the steel ring steps. The fight returned to the ring with Brock squeezing RVD around the ribs. RVD broke away and soon after leveled Brock with a kick to the temple. RVD followed with the Rolling Th under but Brock managed to kick out after a two-count. The fight looked like it might be over soon as RVD went for a Five-Star Frog Splash. Brock, though, hoisted Van Dam onto h is shoulder and went for the F-5, but RVD whirled Brock into a DDT and followed with the Frog Splash! He went for the pin, but Heyman pulled the ref from the ring! The ref immed iately called for a disqualification and soon traded blows with Heyman! After, RVD leapt onto Brock from the top rope and then threatened to hit the Van Terminator! Heyman grabb ed RVD's leg and Brock picked up the champ and this time connected with the F-5 onto a s teel chair! Winner: RVD

Match 6: Booker T vs the Big Show Booker T faced the Big Show one-on-one next. Show withstood Booker T's kicks and punches and slapped Booker into the corner. After being thrown from the ring, Booker picked up a chair at ringside, but Big Show punched it back into Booker's face. Booker tried to get back into the game by choking Show with a camera cable at ringside. Booker smashed a TV monitor from the Sp anish announcers' position into Show's skull, then delivered a scissors kick that put bo th men through the table! Booker crawled back into the ring and Big Show staggered in mo ments later. Show grabbed Booker's throat but was met by a low blow and a kick to the fa ce. Booker climbed the top rope and nailed a somersaulting leg drop to get the pin! Winn er: Booker T

Announcement: Triple H entered the ring to a thunderous ovation

as fans hoped to learn where The Game would end up competing. Before he could speak, Eri c Bishoff stopped The Game to apologize for getting involved in his personal business. I f Triple H signed with RAW, Bischoff promised his personal life would never come into pl ay again. Bischoff said he's spent the past two years networking in Hollywood. He said e veryone was looking for the next breakout WWE Superstar, and they were all talking about Triple H. Bischoff guaranteed that if Triple H signed with RAW, he'd be getting top oppo rtunities coming his way. Stephanie McMahon stepped out to issue her own pitch. She said that because of her personal history with Triple H, the two of them know each other very well. She said the two of them were once unstoppable and they can be again. Bischoff cut her off and begged her to stop. Stephanie cited that Triple H once told her how Bischoff said Triple H had no talent and no charisma. Bischoff said he was young at the time and didn't know what he had, but he still has a lot more experience that Stephanie. The two continued to bicker back and forth, until Triple H stepped up with his microphone. The G ame said it would be easy to say "screw you" to either one of them. Triple H went to sha ke Bischoff's hand, but pulled it away. He said he would rather go with the devil he kno ws, rather than the one he doesn't know. Before he could go any further, though, Shawn M ichaels came out to shake things up. HBK said the last thing he wanted to do was cause a ny trouble. He didn't want to get involved, but he remembered pledging to bring Triple H to the nWo. HBK said there's nobody in the world that Triple H is better friends with. H BK told his friend to imagine the two back together again, making Bischoff's life a livi ng hell. Triple H said that was a tempting offer. He then turned and hugged HBK, making official his switch to RAW! Triple H and HBK left, and Bischoff gloated over his victor y. Bischoff said the difference between the two of them is that he's got testicles and s he doesn't. Stephanie whacked Bischoff on the side of the head and left!

Matc h 7: Tag Team Championship Match Christian and Lance Storm vs Hollywood Hogan and Edge T he match started with loud "USA" chants and with Hogan shoving Christian through the rop es and out of the ring. The Canadians took over from there. But Edge scored a kick to Ch ristian's head and planted a facebuster on Storm to get the tag to Hogan. Hogan began to Hulk up and soon caught Christian with a big boot and a leg drop! Storm broke up the cou nt and Christian tossed Hogan from the ring where Storm superkicked the icon. Edge tagge d in soon after and dropped both opponents. He speared both of them into the corner turn buckles, but missed a spear on Strom and hit the ref hard instead. Edge nailed a DDT, bu t the ref was down and could not count. Test raced down and took down Hogan then leveled Edge with a boot. Storm tried to get the pin, but Edge kicked out after two. Riksihi spr inted in to fend off Test, allowing Edge to recover and spear Storm. Christian distracte d the ref, though, and Y2J dashed in and clocked Edge with the Tag Team Championship! St orm rolled over and got the pinfall to win the title! Winners and New Tag Team Champion s: Christian and Lance Storm

Match 8: WWE Undisputed Championship Triple Thre at Match. The Rock vs Kurt Angle and the Undertaker Three of WWE's most successful super stars lined up against each other in a Triple Threat Match with the Undisputed Champions hip hanging in the balance. Taker and The Rock got face to face with Kurt Angle begging for some attention off to the side. He got attention in the form of a beat down form the two other men. Soon after, Taker spilled out of the ring and The Rock brawled with Angl e. Angle gave a series of suplexes that took down Rock, but the Great One countered with a DDT that managed a two-count. The fight continued outside the ring with Taker coming t o life and clotheslining Angle and repeatedly smacking The Rock. Taker and Rock got into it back into the ring, and Taker dropped The Rock with a sidewalk slam to get a two-coun t. Rock rebounded, grabbed Taker by the throat and chokeslammed him! Angle broke up the pin attempt that likely would have given The Rock the title. The Rock retaliated by latc hing on the ankle lock to Kurt Angle. Angle reversed the move and Rock Bottomed the Peop le's Champion. Soon after, The Rock disposed of Angle and hit the People's Elbow on the Undertaker. Angle tried to take advantage by disabling the Great One outside the ring an d covering Taker, who kicked out after a two count. Outside the ring, Rock took a big sw ig from a nearby water bottle and spewed the liquid into Taker's face to blind the champ ion. Taker didn't stay disabled for long, and managed to overpower Rock and turn his att ention to Angle. Taker landed a guillotine leg drop onto Angle, laying on the ring apro n. The Rock picked himself up just in time to break up a pin attempt on Kurt Angle. Take r nailed Rock with a DDT and set him up for a chokeslam. ANgle tried sneaking up with a steel chair, but Taker caught on to that tomfoolery and smacked it out of his hands. The referee got caught in the ensuing fire and didn't see Angle knock Taker silly with a ste el chair. Angle went to cover Taker as The Rock lay prone, but the Dead Man somehow got his shoulder up. Angle tried to pin Rock, but he too kicked out. The Rock got up and lan ded Angle in the sharpshooter! Angle looked like he was about to tap, but Taker kicked T he Rock out of the submission hold. Taker picked Rock up and crashed him with the Last R ide. While the Dead Man covered him for the win, Angle raced in and picked Taker up in t he ankle lock! Taker went delirious with pain, but managed to counter. He picked Angle u

```
p for the last ride, but Angle put on a triangle choke! It looked like Taker was about t
o pass out, but The Rock broke Angle's hold only to find himself caught in the ankle loc
k. Rock got out of the hold and watched Taker chokeslam Angle. Rocky hit the Rock Botto
m, but Taker refused to go down and kicked out. Angle whirled Taker up into the Angle Sl
am but was Rock Bottomed by the Great One and pinned! Winner and New WWE Champion: The R
ock<br /><br />~Finally there is a decent PPV! Lately the PPV weren't very good, but thi
s one was a winner. I give this PPV a A-<br /><br />

      sentiment  word_count
31481  positive        2470
```

In [12]:
```
#Restoring the displaying option
pd.reset_option('^display.', silent=True)
```

In [13]:
```
#Plotting the word distribution of Positive and Negative reviews
fig,axis= plt.subplots(figsize=(10,6))
axis.hist(df[df['sentiment']=='positive']['word_count'], label='Positive', color='green'
axis.legend(loc='upper right')
axis.hist(df[df['sentiment']=='negative']['word_count'], label='Negative', color='red',
axis.legend(loc='upper right')
fig.suptitle("Number of words in Postive and Negative reviews")
plt.show()
```

In [14]:
```
#Replacing positive and negative sentiments with numerical values for processing
df['sentiment'].replace("positive",1,inplace=True)
df['sentiment'].replace("negative",0,inplace=True)
```

In [15]:
```
df.head()
```

Out[15]:

	review	sentiment	word_count
0	One of the other reviewers has mentioned that ...	1	307
1	A wonderful little production. The...	1	162
2	I thought this was a wonderful way to spend ti...	1	166

3	Basically there's a family where a little boy ...	0	138
4	Petter Mattei's "Love in the Time of Money" is...	1	230

Data Processing

In [16]:
```
#Cleaning the data from mentions, punctuations, urls and tokenising the text to remove s
def text_process(text):
    text= text.lower()
    text= re.sub('<br />','', text)
    text= re.sub(r'https\S+|www\S+|http\S+','',text, flags= re.MULTILINE)
    text=re.sub(r'\@w+|\#','',text)
    text=re.sub(r'[^\w\s]','',text)
    text_tokens = word_tokenize(text)
    filter_text= [w for w in text_tokens if not w in stopwords]
    return(" ".join(filter_text))
```

In [17]:
```
# Applying the function on the dataframe
df.review = df['review'].apply(text_process)
df
```

Out[17]:

	review	sentiment	word_count
0	one reviewers mentioned watching 1 oz episode ...	1	307
1	wonderful little production filming technique ...	1	162
2	thought wonderful way spend time hot summer we...	1	166
3	basically theres family little boy jake thinks...	0	138
4	petter matteis love time money visually stunni...	1	230
...	...	...	...
49995	thought movie right good job wasnt creative or...	1	194
49996	bad plot bad dialogue bad acting idiotic direc...	0	112
49997	catholic taught parochial elementary schools n...	0	230
49998	im going disagree previous comment side maltin...	0	212
49999	one expects star trek movies high art fans exp...	0	129

50000 rows × 3 columns

In [18]:
```
#Finding the duplicates in the data
count_dupli = df.duplicated().sum()
print("The number of duplicates are:", count_dupli)
```

```
The number of duplicates are: 421
```

In [19]:
```
#Removing the duplicates
df=df.drop_duplicates('review')
```

In [20]:
```
#Defining stemming operation to change various forms of the same word
ps= PorterStemmer()
def stem(data):
    text = [ps.stem(word) for word in data.split()]
    #return " ".join(text)
    return data
```

In [21]:
```
#Applying stemming operation to the dataset
df.review = df['review'].apply(lambda x: stem(x))
```

```
C:\Users\Nikhil R Krishnan\AppData\Local\Programs\Python\Python310\lib\site-packages\pan
das\core\generic.py:5516: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
guide/indexing.html#returning-a-view-versus-a-copy
  self[name] = value
```

In [22]:
```
#Finding the first 5 rows of the dataset
df.head()
```

Out[22]:

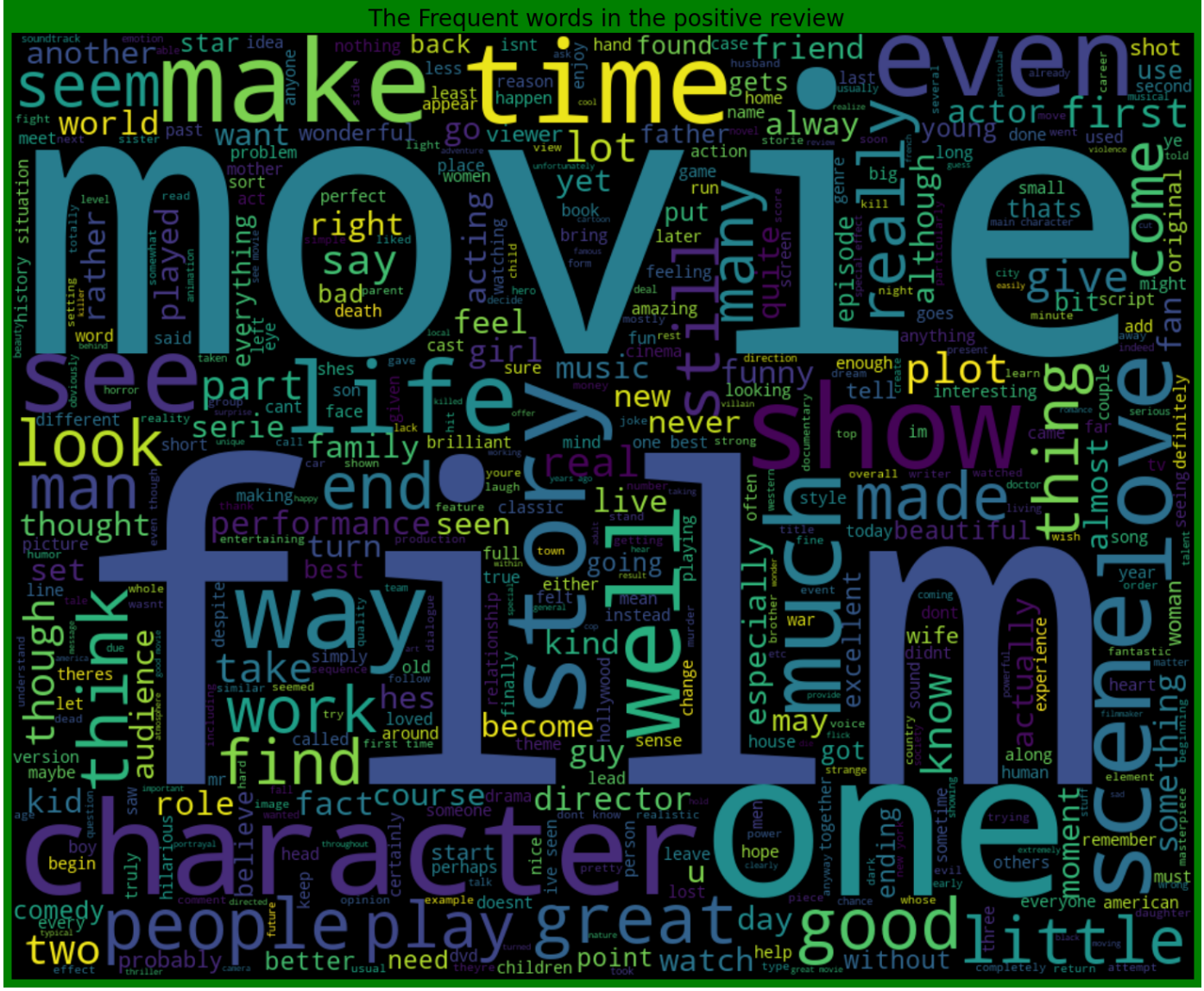
	review	sentiment	word_count
0	one reviewers mentioned watching 1 oz episode ...	1	307
1	wonderful little production filming technique ...	1	162
2	thought wonderful way spend time hot summer we...	1	166
3	basically theres family little boy jake thinks...	0	138
4	petter matteis love time money visually stunni...	1	230

In [23]:
```
#Printing word cloud for postivie reviews
pos_review= df[df.sentiment==1]
pos_review.head()
```

Out[23]:

	review	sentiment	word_count
0	one reviewers mentioned watching 1 oz episode ...	1	307
1	wonderful little production filming technique ...	1	162
2	thought wonderful way spend time hot summer we...	1	166
4	petter matteis love time money visually stunni...	1	230
5	probably alltime favorite movie story selfless...	1	119

In [24]:
```
#Plotting the word cloud
text= ' '.join([word for word in pos_review['review']])
plt.figure(figsize=(20,15), facecolor= 'green')
wordcloud= WordCloud(max_words = 500, width = 1000, height= 800).generate(text)
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.title(" The Frequent words in the positive review", fontsize= 20)
plt.show()
```



In [25]:
```
# The number of times the word is used in these reviews
from collections import Counter
c = Counter()
for t in pos_review['review'].values:
    for w in text.split():
        c[w] += 1
c.most_common(15)
```

```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
Input In [25], in <cell line: 4>()
      4 for t in pos_review['review'].values:
      5     for w in text.split():
----> 6         c[w] += 1
      7 c.most_common(15)

KeyboardInterrupt:
```

In [ ]:
```
pos_words = pd.DataFrame(c.most_common(15))
pos_words.columns = ['word','count']
pos_words.head()
```

In [ ]:
```
px.bar(pos_words, x='count', y='word', title= 'Common words in positive reviews',color=
```

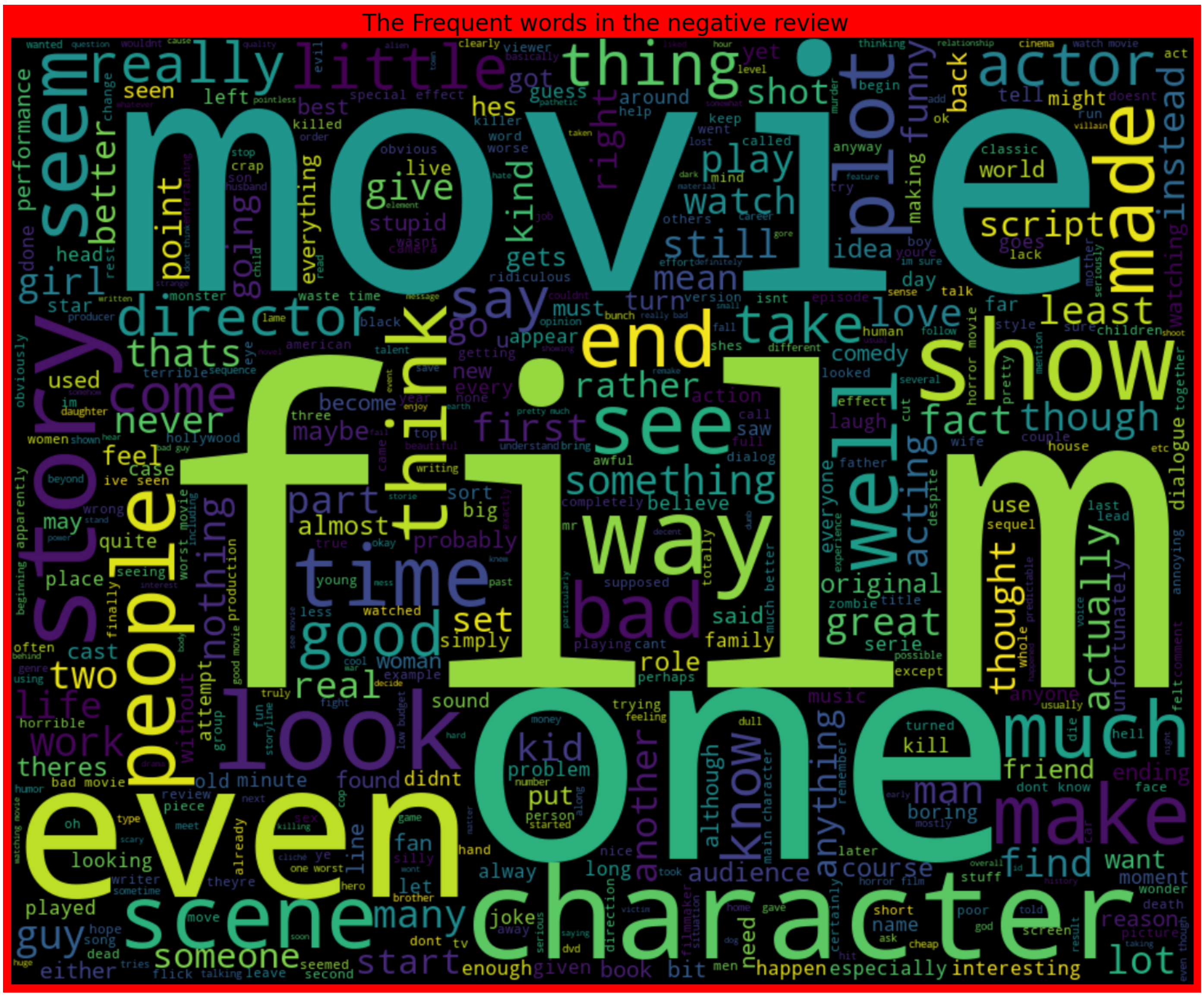
In [26]:
```
#Printing word cloud for negative reviews
neg_review= df[df.sentiment==0]
neg_review.head()
```

Out[26]:

	review	sentiment	word_count
3	basically theres family little boy jake thinks...	0	138
7	show amazing fresh innovative idea 70s first a...	0	174
8	encouraged positive comments film looking forw...	0	130
10	phil alien one quirky films humour based aroun...	0	96
11	saw movie 12 came recall scariest scene big bi...	0	180

In [27]:

```
#Plotting the word cloud
text= ' '.join([word for word in neg_review['review']])
plt.figure(figsize=(20,15), facecolor= 'red')
wordcloud= WordCloud(max_words = 500, width = 1000, height= 800).generate(text)
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.title(" The Frequent words in the negative review", fontsize= 20)
plt.show()
```



In [ ]:

```
# The number of times the word is used in these reviews
from collections import Counter
c = Counter()
for t in neg_review['review'].values:
    for w in text.split():
        c[w] += 1
c.most_common(15)
```

In [ ]:
```
neg_words = pd.DataFrame(c.most_common(15))
neg_words.columns = ['word','count']
neg_words.head()
```

In [ ]:
```
px.bar(neg_words, x='count', y='word', title= 'Common words in positive reviews',color=
```

In [28]:
```
X = df['review']
Y= df['sentiment']
```

In [29]:
```
vect = TfidfVectorizer()
X = vect.fit_transform(df['review'])
```

In [30]:
```
x_train,x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_state=4
```

In [31]:
```
print("The shape of x_train: ",(x_train.shape))
print("The shape of y_train: ",(y_train.shape))
print("The shape of x_test: ",(x_test.shape))
print("The shape of y_test: ",(y_test.shape))
```

```
The shape of x_train:  (34704, 221707)
The shape of y_train:  (34704,)
The shape of x_test:  (14874, 221707)
The shape of y_test:  (14874,)
```

Machine Learning model building

In [56]:
```
#importing Libraries to do the model building
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import warnings
warnings.filterwarnings('ignore')
```

In [33]:
```
logreg= LogisticRegression()
logreg.fit(x_train,y_train)
logreg_predict = logreg.predict(x_test)
logreg_accuracy = accuracy_score(logreg_predict, y_test)
print("Test accuracy: {:.2f}%".format(logreg_accuracy*100))
```

```
Test accuracy: 89.00%
```

In [34]:
```
print(confusion_matrix(y_test, logreg_predict))
print("\n")
print(classification_report(y_test,logreg_predict))
```

```
[[6453  908]
 [ 728 6785]]


              precision    recall  f1-score   support

           0       0.90      0.88      0.89      7361
           1       0.88      0.90      0.89      7513

    accuracy                           0.89     14874
   macro avg       0.89      0.89      0.89     14874
weighted avg       0.89      0.89      0.89     14874
```

In [35]:
```
mnb= MultinomialNB()
```

```
mnb.fit(x_train,y_train)
mnb_predict = mnb.predict(x_test)
mnb_accuracy = accuracy_score(mnb_predict, y_test)
print("Test accuracy: {:.2f}%".format(mnb_accuracy*100))
```

```
Test accuracy: 86.44%
```

In [36]:
```
print(confusion_matrix(y_test, mnb_predict))
print("\n")
print(classification_report(y_test, mnb_predict))
```

```
[[6418  943]
 [1074 6439]]


              precision    recall  f1-score   support

           0       0.86      0.87      0.86      7361
           1       0.87      0.86      0.86      7513

    accuracy                           0.86     14874
   macro avg       0.86      0.86      0.86     14874
weighted avg       0.86      0.86      0.86     14874
```

In [37]:
```
svc= LinearSVC()
svc.fit(x_train,y_train)
svc_predict = svc.predict(x_test)
svc_accuracy = accuracy_score(svc_predict, y_test)
print("Test accuracy: {:.2f}%".format(svc_accuracy*100))
```

```
Test accuracy: 89.22%
```

In [38]:
```
print(confusion_matrix(y_test, svc_predict))
print("\n")
print(classification_report(y_test, svc_predict))
```

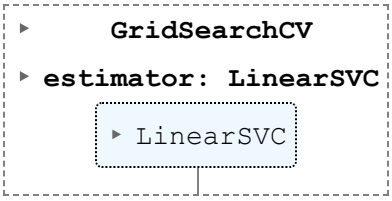
```
[[6504  857]
 [ 747 6766]]


              precision    recall  f1-score   support

           0       0.90      0.88      0.89      7361
           1       0.89      0.90      0.89      7513

    accuracy                           0.89     14874
   macro avg       0.89      0.89      0.89     14874
weighted avg       0.89      0.89      0.89     14874
```

In [39]:
```
from sklearn.model_selection import GridSearchCV
para_grid = {'C':[0.1,1,10,100], 'loss':['hinge', 'squared_hinge']}
grid= GridSearchCV(svc, para_grid, refit = True, verbose = 3)
grid.fit(x_train, y_train)
```

```
Fitting 5 folds for each of 8 candidates, totalling 40 fits
[CV 1/5] END ................C=0.1, loss=hinge;, score=0.872 total time=   0.1s
[CV 2/5] END ................C=0.1, loss=hinge;, score=0.875 total time=   0.1s
[CV 3/5] END ................C=0.1, loss=hinge;, score=0.871 total time=   0.2s
[CV 4/5] END ................C=0.1, loss=hinge;, score=0.878 total time=   0.1s
[CV 5/5] END ................C=0.1, loss=hinge;, score=0.874 total time=   0.3s
[CV 1/5] END .......C=0.1, loss=squared_hinge;, score=0.892 total time=   0.2s
[CV 2/5] END .......C=0.1, loss=squared_hinge;, score=0.895 total time=   0.3s
[CV 3/5] END .......C=0.1, loss=squared_hinge;, score=0.888 total time=   0.1s
[CV 4/5] END .......C=0.1, loss=squared_hinge;, score=0.894 total time=   0.3s
[CV 5/5] END .......C=0.1, loss=squared_hinge;, score=0.890 total time=   0.1s
```

```
[CV 1/5] END ..................C=1, loss=hinge;, score=0.896 total time=   0.5s
[CV 2/5] END ..................C=1, loss=hinge;, score=0.894 total time=   1.4s
[CV 3/5] END ..................C=1, loss=hinge;, score=0.893 total time=   0.4s
[CV 4/5] END ..................C=1, loss=hinge;, score=0.894 total time=   0.5s
[CV 5/5] END ..................C=1, loss=hinge;, score=0.894 total time=   0.3s
[CV 1/5] END ..........C=1, loss=squared_hinge;, score=0.892 total time=   0.4s
[CV 2/5] END ..........C=1, loss=squared_hinge;, score=0.895 total time=   0.7s
[CV 3/5] END ..........C=1, loss=squared_hinge;, score=0.889 total time=   0.4s
[CV 4/5] END ..........C=1, loss=squared_hinge;, score=0.896 total time=   0.3s
[CV 5/5] END ..........C=1, loss=squared_hinge;, score=0.894 total time=   0.3s
[CV 1/5] END .................C=10, loss=hinge;, score=0.875 total time=   2.0s
[CV 2/5] END .................C=10, loss=hinge;, score=0.882 total time=   7.6s
[CV 3/5] END .................C=10, loss=hinge;, score=0.875 total time=   5.4s
[CV 4/5] END .................C=10, loss=hinge;, score=0.881 total time=   2.2s
[CV 5/5] END .................C=10, loss=hinge;, score=0.878 total time=   4.4s
[CV 1/5] END ..........C=10, loss=squared_hinge;, score=0.881 total time=   0.9s
[CV 2/5] END ..........C=10, loss=squared_hinge;, score=0.885 total time=   1.2s
[CV 3/5] END ..........C=10, loss=squared_hinge;, score=0.879 total time=   1.5s
[CV 4/5] END ..........C=10, loss=squared_hinge;, score=0.885 total time=   1.2s
[CV 5/5] END ..........C=10, loss=squared_hinge;, score=0.883 total time=   1.4s
[CV 1/5] END ................C=100, loss=hinge;, score=0.876 total time=   1.5s
[CV 2/5] END ................C=100, loss=hinge;, score=0.881 total time=   5.8s
[CV 3/5] END ................C=100, loss=hinge;, score=0.874 total time=   6.5s
[CV 4/5] END ................C=100, loss=hinge;, score=0.880 total time=   3.1s
[CV 5/5] END ................C=100, loss=hinge;, score=0.878 total time=   6.9s
[CV 1/5] END .........C=100, loss=squared_hinge;, score=0.877 total time=   1.3s
[CV 2/5] END .........C=100, loss=squared_hinge;, score=0.882 total time=   1.8s
[CV 3/5] END .........C=100, loss=squared_hinge;, score=0.875 total time=   4.4s
[CV 4/5] END .........C=100, loss=squared_hinge;, score=0.881 total time=   4.0s
[CV 5/5] END .........C=100, loss=squared_hinge;, score=0.878 total time=   4.8s
```

Out[39]:



In [40]:
```
print("best cross validation score:  {:.2f}".format(grid.best_score_))
print("best parameters: ", grid.best_params_)
```

```
best cross validation score:  0.89
best parameters:  {'C': 1, 'loss': 'hinge'}
```

In [41]:
```
svc= LinearSVC(C = 1, loss = 'hinge')
svc.fit(x_train,y_train)
svc_predict = svc.predict(x_test)
svc_accuracy = accuracy_score(svc_predict, y_test)
print("Test accuracy: {:.2f}%".format(svc_accuracy*100))
```

```
Test accuracy: 89.41%
```

In [42]:
```
print(confusion_matrix(y_test, svc_predict))
print("\n")
print(classification_report(y_test, svc_predict))
```

```
[[6511  850]
 [ 725 6788]]


              precision    recall  f1-score   support

           0       0.90      0.88      0.89      7361
           1       0.89      0.90      0.90      7513

    accuracy                           0.89     14874
```

```
   macro avg       0.89      0.89      0.89     14874
weighted avg       0.89      0.89      0.89     14874
```

Importing Tweet to calculate the sentiments

In [43]:
```
#importing the csv file of tweet for movie reviews
df_tweet = pd.read_csv('movie_reviews_tweets.csv')
```

Exploratory Analysis

In [44]:
```
df_tweet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200004 entries, 0 to 200003
Data columns (total 6 columns):
 #   Column       Non-Null Count   Dtype 
---  ------       --------------   ----- 
 0   Unnamed: 0   200004 non-null  int64 
 1   Datetime     200004 non-null  object
 2   Tweet Id     200004 non-null  int64 
 3   Text         200004 non-null  object
 4   Username     200004 non-null  object
 5   Location     3567 non-null    object
dtypes: int64(2), object(4)
memory usage: 9.2+ MB
```

In [45]:
```
# Finding first five rows of data
df_tweet.head()
```

Out[45]:

	Unnamed: 0	Datetime	Tweet Id	Text	Username	Location
0	0	2022-09-05 20:24:52+00:00	1566885092483547138	Rewatched @MichaelJaiWhite at his best in "Blo...	dravencage	NaN
1	1	2022-09-05 20:00:12+00:00	1566878885903536130	Mark Hamill Is Still Bitter 'Star Wars' Sequel...	BoxReview_	NaN
2	2	2022-09-05 20:00:11+00:00	1566878884137783298	'The Mandalorian' Star Teases Gideon's Big Pla...	BoxReview_	NaN
3	3	2022-09-05 20:00:00+00:00	1566878837564121089	Hello Movie Lovers! (Announcement)\n\nTwo Witc...	bobsmoviereview	NaN
4	4	2022-09-05 19:46:04+00:00	1566875330274627584	New episode featuring @letstalkhorrorchannel i...	EveryMovieEver_	NaN

Data cleaning

In [46]:
```
# Removing the unwanted columns
df_tweet.drop(['Unnamed: 0','Datetime','Tweet Id','Username','Location'], axis=1, inplac
```

In [47]:
```
# Printing the new dataframe
df_tweet.head()
```

Out[47]:

	Text

0	Rewatched @MichaelJaiWhite at his best in "Blo...
1	Mark Hamill Is Still Bitter 'Star Wars' Sequel...
2	'The Mandalorian' Star Teases Gideon's Big Pla...
3	Hello Movie Lovers! (Announcement)\n\nTwo Witc...
4	New episode featuring @letstalkhorrorchannel i...

In [48]:
```
#Cleaning the data from mentions, punctuations, urls and tokenising the text to remove s
def CleanText(text):
    text= text.lower()
    text= re.sub('<br />','', text)
    text= re.sub('&amp|amp','', text)
    text= re.sub(r'https\S+|www\S+|http\S+','',text, flags= re.MULTILINE)
    text=re.sub(r'\@w+|\#','',text)
    text=re.sub(r'[^\w\s]','',text)
    text_tokens = word_tokenize(text)
    filter_text= [w for w in text_tokens if not w in stopwords]
    return(" ".join(filter_text))

#Cleaning the texts in the tweets
df_tweet.Text = df_tweet['Text'].apply(CleanText)

#First five rows of clean data
df_tweet
```

Out[48]:

	Text
0	rewatched michaeljaiwhite best blood bone blas...
1	mark hamill still bitter star wars sequels did...
2	mandalorian star teases gideons big plans empi...
3	hello movie lovers announcement two witches bl...
4	new episode featuring letstalkhorrorchannel li...
...	...
199999	movie review featuring nischay sapphire shivam...
200000	thegentlemen jumanjithenextlevel sunday movie ...
200001	read post film filmreview films horror movie m...
200002	officially mind blown 3 students review bongjo...
200003	spending sundaymorning check philly arts cultu...

200004 rows × 1 columns

In [49]:
```
df_tweet.isnull().sum()
```

Out[49]:
```
Text    0
dtype: int64
```

In [50]:
```
#Removing Duplicates
df_tweet.drop_duplicates(subset='Text',inplace=True, ignore_index= True)
df_tweet
```

Out[50]:

	Text
0	rewatched michaeljaiwhite best blood bone blas...

1	mark hamill still bitter star wars sequels did...
2	mandalorian star teases gideons big plans empi...
3	hello movie lovers announcement two witches bl...
4	new episode featuring letstalkhorrorchannel li...
...	...
149245	film review fistful dollars eastwood ultra coo...
149246	review twomeninmanhattan letterboxd criterionc...
149247	vids dig 224 midnights edge phantom retrospect...
149248	movie review featuring nischay sapphire shivam...
149249	thegentlemen jumanjithenextlevel sunday movie ...

149250 rows × 1 columns

Labelling the sentiments of the data using Linear SVC (Model found to have best accuracy

In [60]:
```
train_vectors = X
train_label = Y
test_vectors= vect.transform(df_tweet['Text'])
```

In [61]:
```
classifier_linear = svm.LinearSVC()
classifier_linear.fit(train_vectors, train_label)
review_vector = vect.transform(df_tweet['Text']) # vectorizing
df_tweet['label']=  classifier_linear.predict(review_vector)
```

In [62]:
```
df_tweet
```

Out[62]:

	Text	label
0	rewatched michaeljaiwhite best blood bone blas...	1
1	mark hamill still bitter star wars sequels did...	0
2	mandalorian star teases gideons big plans empi...	0
3	hello movie lovers announcement two witches bl...	0
4	new episode featuring letstalkhorrorchannel li...	1
...	...	...
149245	film review fistful dollars eastwood ultra coo...	1
149246	review twomeninmanhattan letterboxd criterionc...	1
149247	vids dig 224 midnights edge phantom retrospect...	1
149248	movie review featuring nischay sapphire shivam...	1
149249	thegentlemen jumanjithenextlevel sunday movie ...	1

149250 rows × 2 columns

In [64]:
```
# Creating a function to compute the negative and positive analysis
def getAnalys(score):
```

```
    if (score==0):
        return 'Negative'
    else:
        return 'Positive'

df_tweet['Analysis'] = df_tweet['label'].apply(getAnalys)

#Showing the dataframe
df_tweet
```

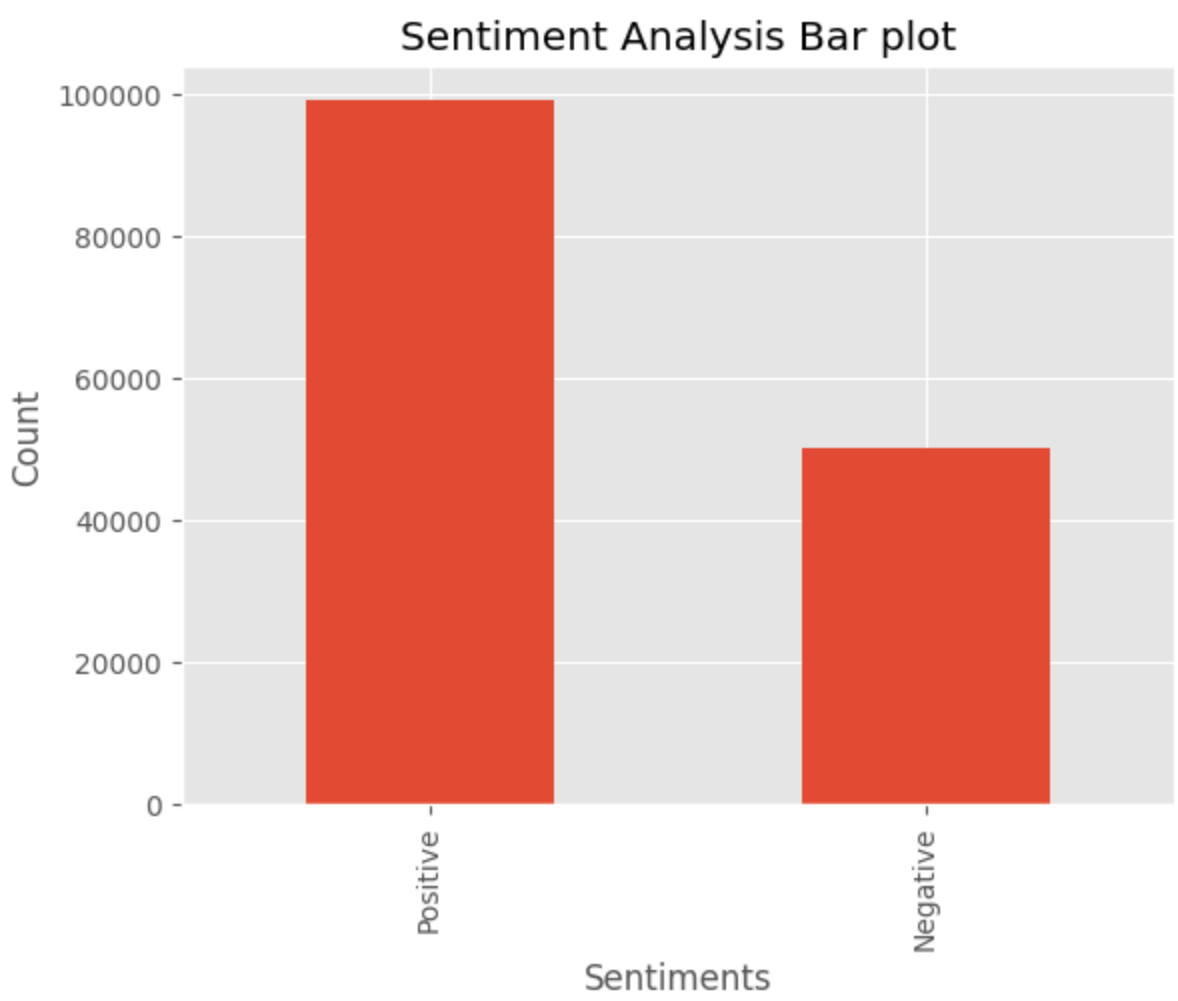
Out[64]:

	Text	label	Analysis
0	rewatched michaeljaiwhite best blood bone blas...	1	Positive
1	mark hamill still bitter star wars sequels did...	0	Negative
2	mandalorian star teases gideons big plans empi...	0	Negative
3	hello movie lovers announcement two witches bl...	0	Negative
4	new episode featuring letstalkhorrorchannel li...	1	Positive
...	...	...	...
149245	film review fistful dollars eastwood ultra coo...	1	Positive
149246	review twomeninmanhattan letterboxd criterionc...	1	Positive
149247	vids dig 224 midnights edge phantom retrospect...	1	Positive
149248	movie review featuring nischay sapphire shivam...	1	Positive
149249	thegentlemen jumanjithenextlevel sunday movie ...	1	Positive

149250 rows × 3 columns

In [66]:

```
# Creating Bar chart to show the count of Positive and Negative sentiments
df_tweet['Analysis'].value_counts().plot(kind='bar')
plt.title('Sentiment Analysis Bar plot')
plt.xlabel('Sentiments')
plt.ylabel('Count')
plt.show()
```

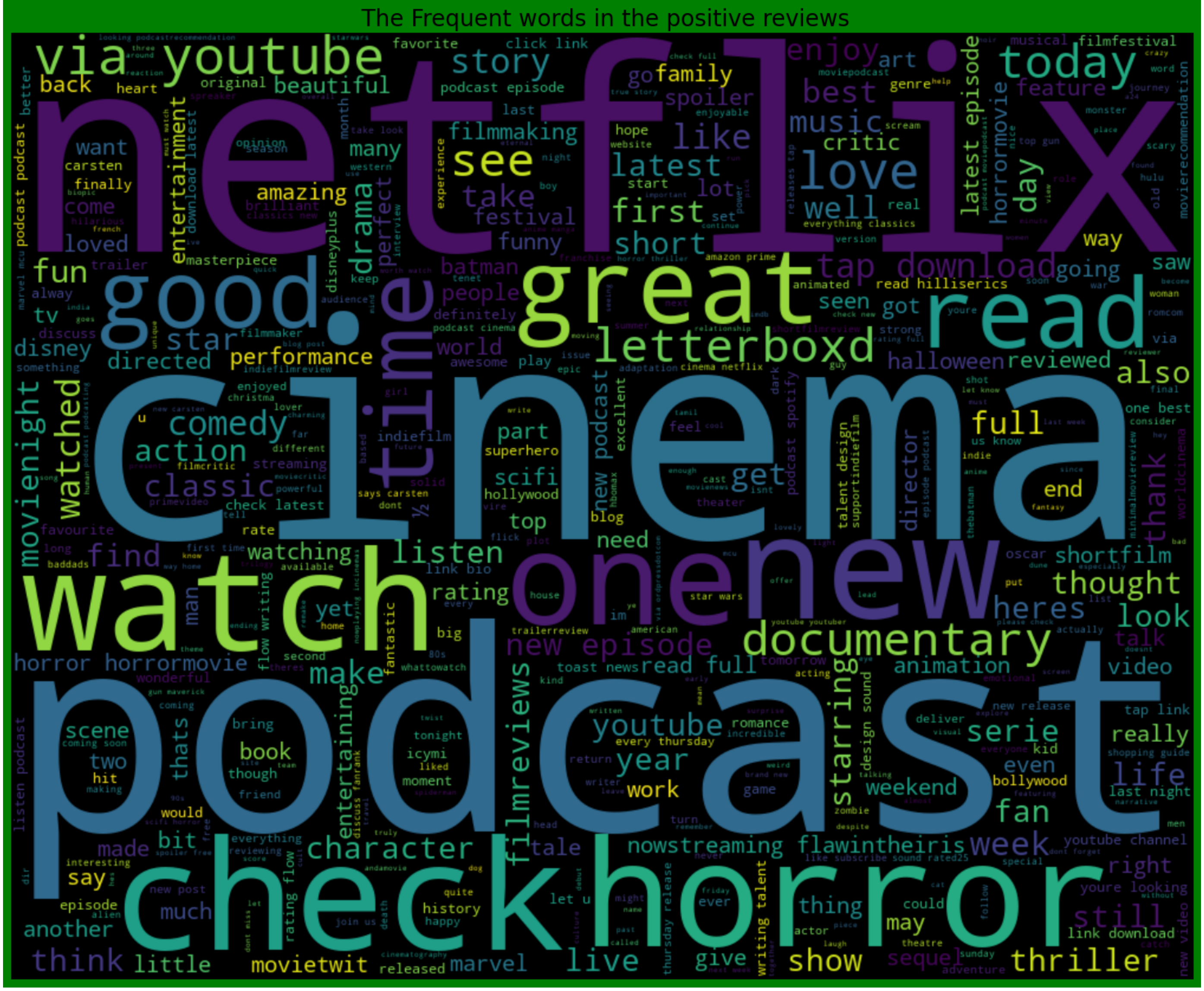


In [74]:
```
#Printing word cloud for postivie reviews
pos_tweet_review= df_tweet[df_tweet.label==1]
pos_tweet_review.head()
```

Out[74]:

	Text	label	Analysis
0	rewatched michaeljaiwhite best blood bone blas...	1	Positive
4	new episode featuring letstalkhorrorchannel li...	1	Positive
5	check buckbreaking movie review tariqnasheed t...	1	Positive
7	words immortal bruce cbell reviews review movi...	1	Positive
13	hello movie lovers announcement paranormal act...	1	Positive

In [78]:
```
#Plotting the word cloud
text= ' '.join([word for word in pos_tweet_review['Text']])
restricted = ['movie','review','moviereview','filmreview','film','moviereviews','films',

plt.figure(figsize=(20,15), facecolor= 'green')
wordcloud= WordCloud(stopwords = restricted, max_words = 500, width = 1000, height= 800)
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.title(" The Frequent words in the positive reviews", fontsize= 20)
plt.show()
```



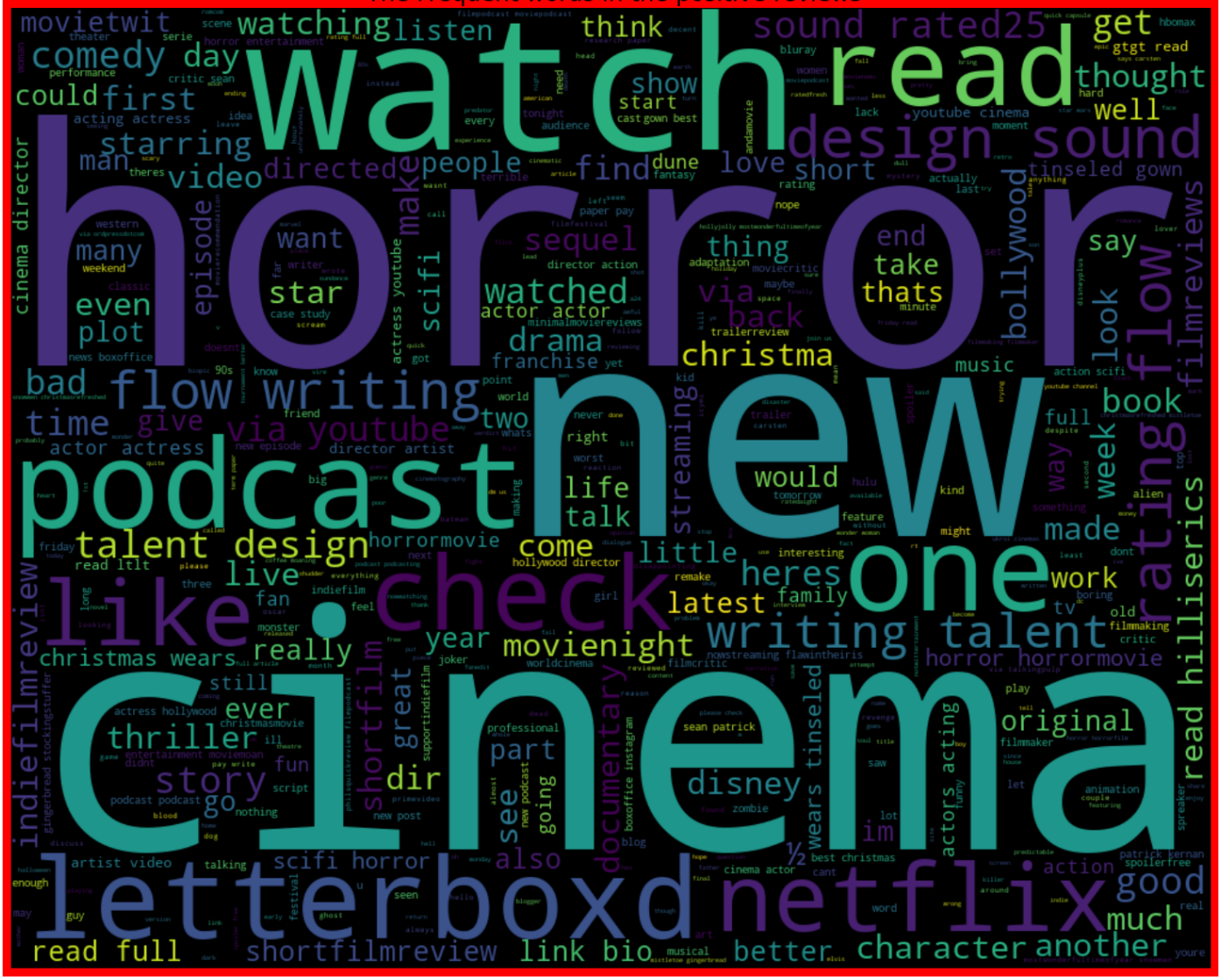
In [75]:
```
#Printing word cloud for negative reviews
neg_tweet_review= df_tweet[df_tweet.label==0]
neg_tweet_review.head()
```

Out[75]:

	Text	label	Analysis
1	mark hamill still bitter star wars sequels did...	0	Negative
2	mandalorian star teases gideons big plans empi...	0	Negative
3	hello movie lovers announcement two witches bl...	0	Negative
6	attack titan getting stage musical adaptation ...	0	Negative
8	musical comedy penelope development disney gtg...	0	Negative

In [77]:
```
#Plotting the word cloud
text= ' '.join([word for word in neg_tweet_review['Text']])
restricted = ['movie','review','moviereview','filmreview','film','moviereviews','films',

plt.figure(figsize=(20,15), facecolor= 'red')
wordcloud= WordCloud(stopwords = restricted, max_words = 500, width = 1000, height= 800)
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis('off')
plt.title(" The Frequent words in the negtive reviews", fontsize= 20)
plt.show()
```

The Frequent words in the positive reviews



Pickling the model to be used in flask app

In [80]:
```
import pickle

# pickling the vectorizer
pickle.dump(vect, open('vectorizer.sav', 'wb'))
# pickling the model
pickle.dump(classifier_linear, open('classifier.sav', 'wb'))

print('Both vectorizer and classifier has been pickled. Check "classifier_flask" to load
```

```
Both vectorizer and classifier has been pickled. Check "classifier_flask" to load and us
e in flask app
```

In [ ]: