1 Social media Sentimental analysis using NLTK-ML-Task-2

In [1]:
```
'''Sentiment analysis, also known as opinion mining, is the process of using␣
↪natural language processing (NLP) techniques
to determine the sentiment or emotional tone expressed in text data.
When applied to social media data, sentiment analysis can provide valuable␣
↪insights into public opinion, customer feedback,
brand perception, and more. Here's a brief overview of sentiment analysis using␣
↪social media data:
'''
```

Out[1]:
```
"Sentiment analysis, also known as opinion mining, is the process of using␣\n↪natural language processing (NLP) techn
iques\nto determine the sentiment or emotional tone expressed in text data.\nWhen applied to social media data, senti
ment analysis can provide valuable␣\n↪insights into public opinion, customer feedback,\nbrand perception, and more. H
ere's a brief overview of sentiment analysis using␣\n↪social media data:\n"
```

In [2]:
```
'''Use adataset of tweets or Facebook posts and perfom sentimental analysis to determine the overall ssentiment of the
```

Out[2]:
```
'Use adataset of tweets or Facebook posts and perfom sentimental analysis to determine the overall ssentiment of the
posts'
```

In [3]:
```
import nltk
```

In [ ]:
```

```

In [4]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
import string
import nltk
from nltk.corpus import stopwords
from nltk import PorterStemmer
import string
import re
from wordcloud import WordCloud
#from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

Read the data

In [5]:
```
file_path='C:\\Users\\Dayakar\\Desktop\\DS  Assigments\\internship 27\\archive (3)\\Tweets.csv'
tweets_df=pd.read_csv(file_path)
tweets_df
```

Out[5]:

airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	us
Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	(U
Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	NaN	(U
Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	(U
Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	NaN	(U
Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	(U
...	...	...	...	...	...	...	...	...	
American	NaN	KristenReenders	NaN	0	@AmericanAir thank you we got on a different f...	NaN	2015-02-22 12:01:01 -0800	NaN	
American	NaN	itsropes	NaN	0	@AmericanAir leaving over 20 minutes Late Flig...	NaN	2015-02-22 11:59:46 -0800	Texas	
American	NaN	sanyabun	NaN	0	@AmericanAir Please bring American Airlines to...	NaN	2015-02-22 11:59:15 -0800	Nigeria,lagos	
American	NaN	SraJackson	NaN	0	@AmericanAir you have my money, you change my ...	NaN	2015-02-22 11:59:02 -0800	New Jersey	(U
American	NaN	daviddtwu	NaN	0	@AmericanAir we have 8 ppl so we need 2 know h...	NaN	2015-02-22 11:58:51 -0800	dallas, TX	

In [6]:
```
tweets_df.size
```

Out[6]:
```
219600
```

In [7]:
```
tweets_df.shape
```

Out[7]:
```
(14640, 15)
```

In [8]:
```
tweets_df.columns
```

Out[8]:
```
Index(['tweet_id', 'airline_sentiment', 'airline_sentiment_confidence',
       'negativereason', 'negativereason_confidence', 'airline',
       'airline_sentiment_gold', 'name', 'negativereason_gold',
       'retweet_count', 'text', 'tweet_coord', 'tweet_created',
       'tweet_location', 'user_timezone'],
      dtype='object')
```

In [9]:
```
tweets_df.dtypes
```

Out[9]:
```
tweet_id                         int64
airline_sentiment               object
airline_sentiment_confidence    float64
negativereason                  object
negativereason_confidence       float64
airline                         object
airline_sentiment_gold          object
name                            object
negativereason_gold             object
retweet_count                    int64
text                            object
tweet_coord                     object
tweet_created                   object
tweet_location                  object
user_timezone                   object
dtype: object
```

In [10]:
```
tweets_df.isnull().sum()
```

Out[10]:
```
tweet_id                        0
airline_sentiment               0
airline_sentiment_confidence    0
negativereason                5462
negativereason_confidence     4118
airline                         0
airline_sentiment_gold        14600
name                            0
negativereason_gold           14608
retweet_count                   0
text                            0
tweet_coord                   13621
tweet_created                   0
tweet_location                4733
user_timezone                 4820
dtype: int64
```

In [11]:
```
tweets_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   tweet_id                      14640 non-null  int64  
 1   airline_sentiment             14640 non-null  object 
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object 
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object 
 6   airline_sentiment_gold        40 non-null     object 
 7   name                          14640 non-null  object 
 8   negativereason_gold           32 non-null     object 
 9   retweet_count                 14640 non-null  int64  
 10  text                          14640 non-null  object 
 11  tweet_coord                   1019 non-null   object 
 12  tweet_created                 14640 non-null  object 
 13  tweet_location                9907 non-null   object 
 14  user_timezone                 9820 non-null   object 
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

In [12]:
```
cols=['negativereason','negativereason_confidence','airline_sentiment_gold']
tweets_df.drop(cols,axis=1,inplace=True)
```

In [13]:
```
tweets_df.isnull().sum()
```

Out[13]:
```
tweet_id                          0
airline_sentiment                 0
airline_sentiment_confidence      0
airline                           0
name                              0
negativereason_gold           14608
retweet_count                     0
text                              0
tweet_coord                   13621
tweet_created                     0
tweet_location                 4733
user_timezone                  4820
dtype: int64
```

In [14]:
```
cols=['negativereason_gold','tweet_coord','tweet_location','user_timezone']
tweets_df.drop(cols,axis=1,inplace=True)
```

In [15]:
```
tweets_df.isnull().sum()
```

Out[15]:
```
tweet_id                      0
airline_sentiment             0
airline_sentiment_confidence  0
airline                       0
name                          0
retweet_count                 0
text                          0
tweet_created                 0
dtype: int64
```

In [16]:
```
tweets_df.shape
```

Out[16]:
```
(14640, 8)
```

In [17]:
```
tweets_df=df = tweets_df[['airline_sentiment','text']]
tweets_df
```

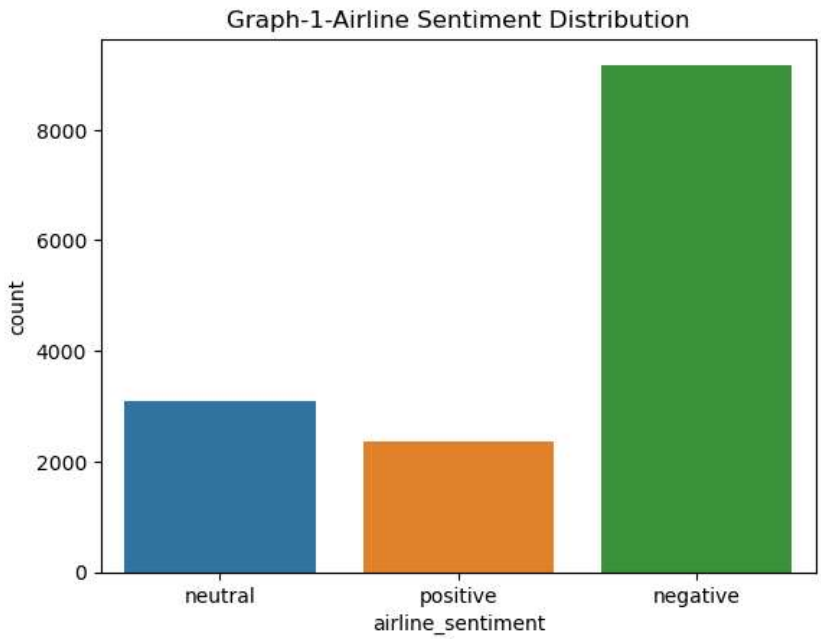
Out[17]:

	airline_sentiment	text
0	neutral	@VirginAmerica What @dhepburn said.
1	positive	@VirginAmerica plus you've added commercials t...
2	neutral	@VirginAmerica I didn't today... Must mean I n...
3	negative	@VirginAmerica it's really aggressive to blast...
4	negative	@VirginAmerica and it's a really big bad thing...
...	...	...
14635	positive	@AmericanAir thank you we got on a different f...
14636	negative	@AmericanAir leaving over 20 minutes Late Flig...
14637	neutral	@AmericanAir Please bring American Airlines to...
14638	negative	@AmericanAir you have my money, you change my ...
14639	neutral	@AmericanAir we have 8 ppl so we need 2 know h...

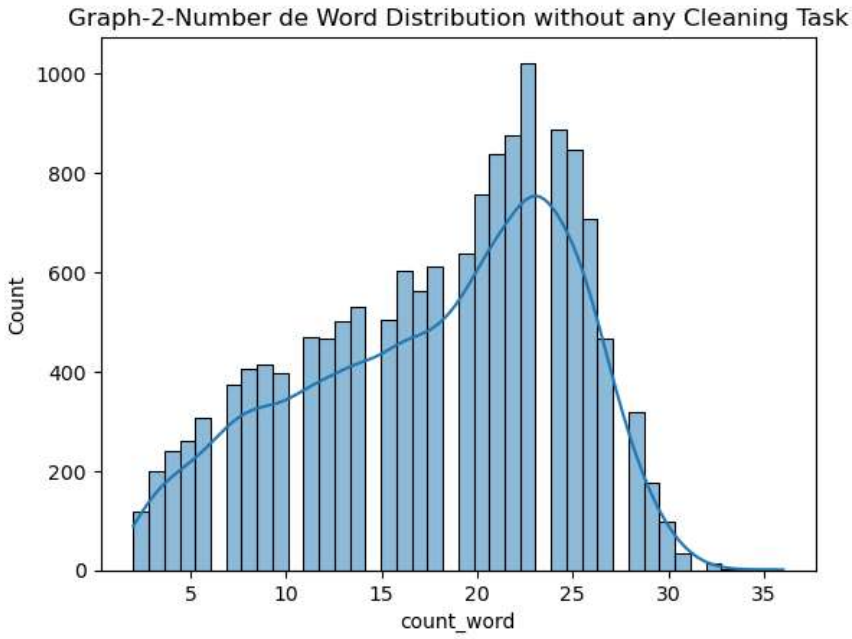
14640 rows × 2 columns

In [18]:
```
sns.countplot(data=tweets_df,x='airline_sentiment')
plt.title('Graph-1-Airline Sentiment Distribution')
```

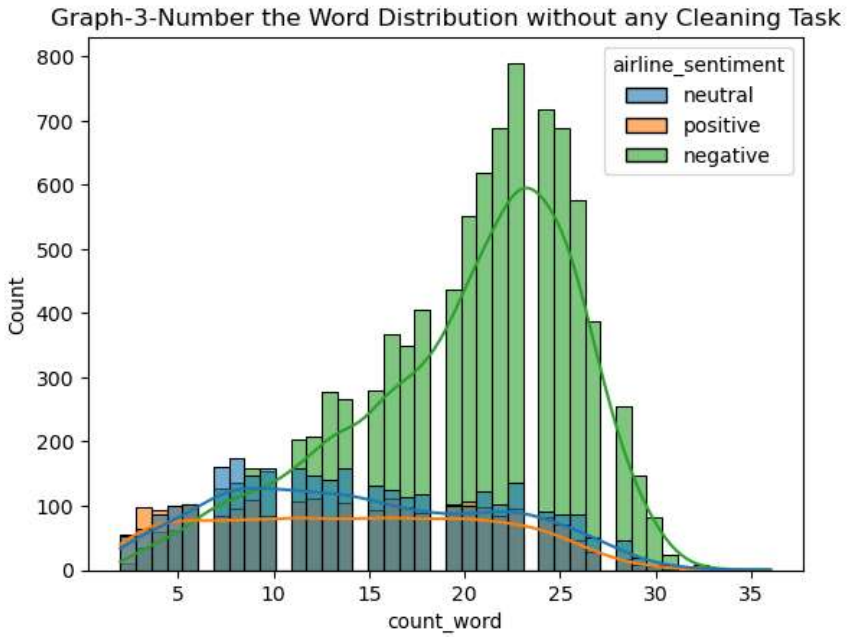
Out[18]:
```
Text(0.5, 1.0, 'Graph-1-Airline Sentiment Distribution')
```



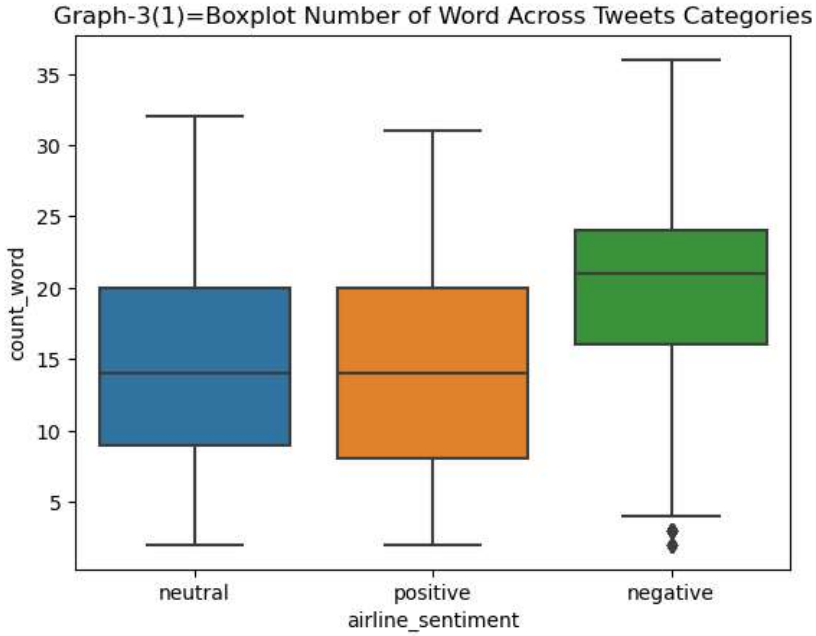
In [19]:
```
tweets_df['count_word'] = tweets_df['text'].apply(lambda x : len(x.split(' ')))
sns.histplot(data = tweets_df , x='count_word',kde=True)
plt.title('Graph-2-Number de Word Distribution without any Cleaning Task')
plt.show()
```



In [20]:
```
sns.histplot(data = tweets_df , x='count_word',hue='airline_sentiment',alpha=0.6,kde=True)
plt.title('Graph-3-Number the Word Distribution without any Cleaning Task')
plt.show()
```



In [21]:
```
sns.boxplot(data = tweets_df , y='count_word',x='airline_sentiment')
plt.title('Graph-3(1)=Boxplot Number of Word Across Tweets Categories')
plt.show()
```



In [22]:
```
df.loc[np.logical_or(df['count_word']>35,df['count_word']<=5),:]
```

Out[22]:

	airline_sentiment	text	count_word
0	neutral	@VirginAmerica What @dhepburn said.	4
14	positive	@VirginAmerica Thanks!	2
18	positive	I ❤️ flying @VirginAmerica. ☺👍	5
46	neutral	@VirginAmerica DREAM http://t.co/oA2dRfAoQ2 h...	5
58	neutral	@VirginAmerica @ladygaga @carrieunderwood - Ca...	5
...	...	...	...
14312	positive	@AmericanAir awesome! Thx	3
14314	negative	@AmericanAir yes, and rebooked incorrectly.	5
14443	neutral	@AmericanAir hi how are you	5
14600	neutral	http://t.co/Elw2sYb8Fu roberts&s=1 @Americ...	3
14630	positive	@AmericanAir Thanks! He is.	4

817 rows × 3 columns

preprocessing the data

In [23]:
```
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
```

In [24]:
```
# punctuation Removal
def remove_punctuation(text):
    return re.sub(r'[^\w\s]','',text)
#stopword removal
def remove_stopwords(text):
    stop_words = set(stopwords.words('english'))
    tokens = word_tokenize(text)
    filter_tokens = [word for word in tokens if word.lower() not in stop_words]
    return " ".join(filter_tokens)
#remove numeric
def remove_numeric(text):
    return re.sub(r'\d+','',text)
#Stemming
def apply_stemming(text):
    stemmer = PorterStemmer()
    tokens = word_tokenize(text)
    stemmed_tokens = [stemmer.stem(word) for word in tokens]
    return " ".join(stemmed_tokens)
def remove_mentions(text):
    return re.sub(r'@\w+','',text)
```

In [25]:
```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
def apply_stemming(text):
    stemmer = PorterStemmer()
    tokens = word_tokenize(text)
    stemmed_tokens = [stemmer.stem(word) for word in tokens]
    return " ".join(stemmed_tokens)
input_text = "walking throw the street, a passenger walked toward me, talking about a walked chicken on the streets"
stemmed_text = apply_stemming(input_text)
print(stemmed_text)
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Dayakar\AppData\Roaming\nltk_data...

walk throw the street , a passeng walk toward me , talking about a walk chicken on the street

[nltk_data]   Package punkt is already up-to-date!
```

In [26]:
```
apply_stemming('walking throw the street , a passenger walked toward me talking about a walked chicken on the streets')
```

Out[26]:
```
'walk throw the street , a passeng walk toward me talk about a walk chicken on the street'
```

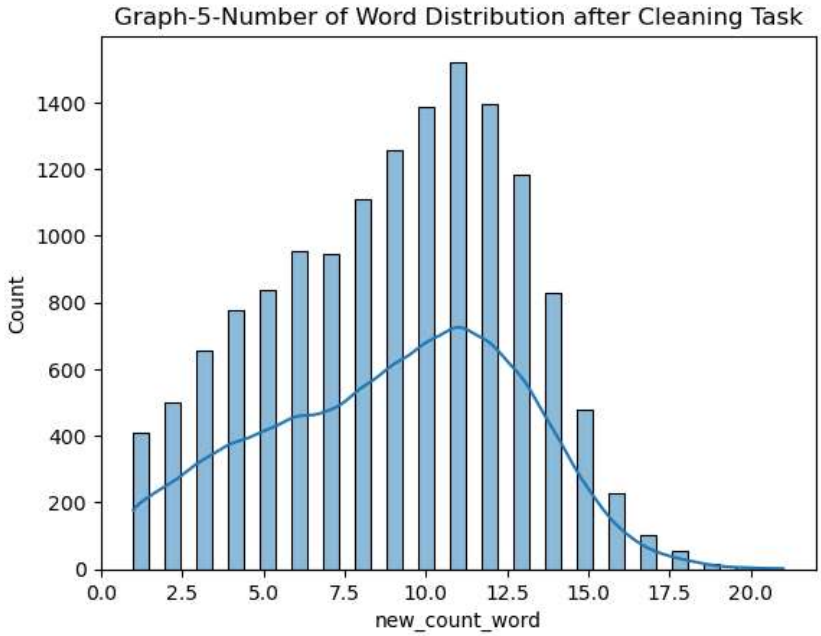
In [27]:
```
def text_preprocessing(text):
    sentence = remove_mentions(text)
    sentence = remove_punctuation(sentence)
    sentence = remove_stopwords(sentence)
    sentence = remove_numeric(sentence)
    sentence = apply_stemming(sentence)
    return sentence
```

In [28]:
```
text_preprocessing('walking throw the street , a passenger walked toward me␣,talking about a walked chicken on the stre
```

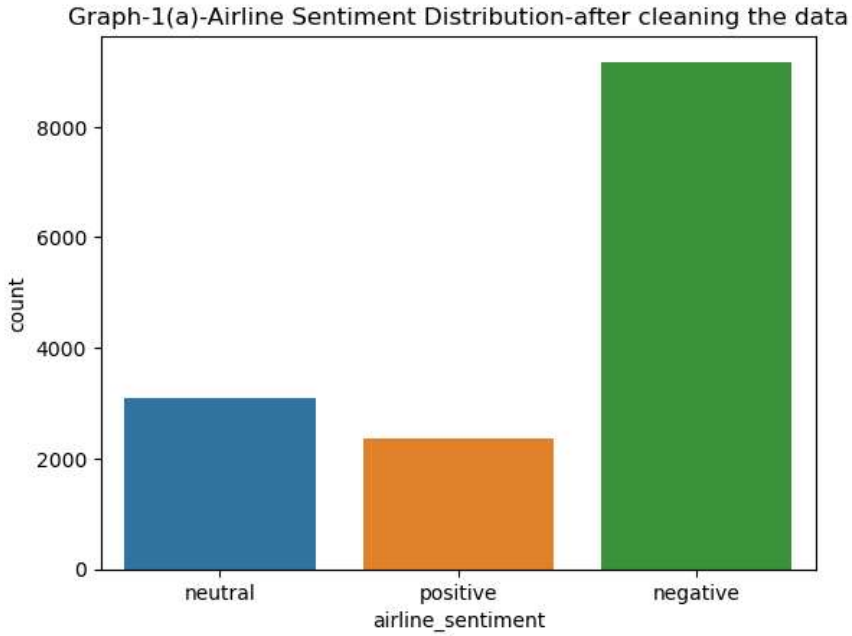
Out[28]:
```
'walk throw street passeng walk toward metalk walk chicken street'
```

In [31]:
```
tweets_df.loc[:,'new_text'] = tweets_df['text'].apply(lambda x : text_preprocessing(x))
```

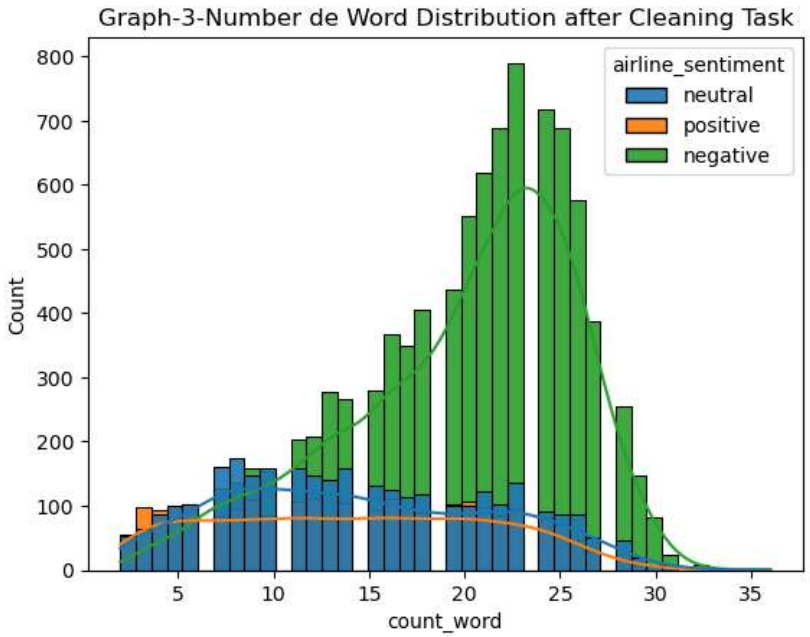
In [32]:
```
tweets_df.loc[:,'new_count_word'] = tweets_df['new_text'].apply(lambda x : len(x.split(' ')))
sns.histplot(data = tweets_df , x='new_count_word',kde=True)
plt.title('Graph-5-Number of Word Distribution after Cleaning Task')
plt.show()
```



In [33]:
```
sns.countplot(data=tweets_df,x='airline_sentiment')
plt.title('Graph-1(a)-Airline Sentiment Distribution-after cleaning the data')
plt.show()
```



In [34]:
```
sns.histplot(data = tweets_df , x='count_word',hue='airline_sentiment',alpha=0.9,kde=True)
plt.title('Graph-3-Number de Word Distribution after Cleaning Task')
plt.show()
```



In [ ]: