NLP Phase 2

Purvaj Desai

A20469336

Primary Dataset

In [13]:
```
from sklearn.metrics import cohen_kappa_score
import pandas as pd
```

1) nyt_stance_0

In [23]:
```
df_nyt_stance_0 = pd.read_csv("C:\\College\\SEM 2\\NLP\\Assignment 2\\nyt_sta
df_nyt_stance_0.isnull().sum()
```

Out[23]:
```
text            0
annotation_64    1
annotation_60    0
annotation_61    0
annotation_63    0
dtype: int64
```

Currently we have one missing value

Replacing Null Value with "missing"

In [24]:
```
df_nyt_stance_0 = df_nyt_stance_0.fillna(value = "missing")
df_nyt_stance_0.isnull().sum()
```

Out[24]:
```
text            0
annotation_64    0
annotation_60    0
annotation_61    0
annotation_63    0
dtype: int64
```

In [57]:

```
nyt_stance_score_0_64_60 = cohen_kappa_score(df_nyt_stance_0.annotation_64, d
nyt_stance_score_0_64_61 = cohen_kappa_score(df_nyt_stance_0.annotation_64,df
nyt_stance_score_0_64_63 = cohen_kappa_score(df_nyt_stance_0.annotation_64,df
nyt_stance_score_0_60_61 = cohen_kappa_score(df_nyt_stance_0.annotation_60,df
nyt_stance_score_0_60_63 = cohen_kappa_score(df_nyt_stance_0.annotation_60,df
nyt_stance_score_0_61_63 = cohen_kappa_score(df_nyt_stance_0.annotation_61,df

avg_nyt_stance_score_0_64 = (nyt_stance_score_0_64_60 + nyt_stance_score_0_64
avg_nyt_stance_score_0_60 = (nyt_stance_score_0_64_60 + nyt_stance_score_0_60
avg_nyt_stance_score_0_61 = (nyt_stance_score_0_60_61 + nyt_stance_score_0_64
avg_nyt_stance_score_0_63 = (nyt_stance_score_0_64_63 + nyt_stance_score_0_60

print("Average Score of Each Annotator for nyt_stance_0 File \n")
print("Average Score of Annotator 64:",avg_nyt_stance_score_0_64)
print("Average Score of Annotator 60:",avg_nyt_stance_score_0_60)
print("Average Score of Annotator 61:",avg_nyt_stance_score_0_61)
print("Average Score of Annotator 63:",avg_nyt_stance_score_0_63)
```

```
Average Score of Each Annotator for nyt_stance_0 File

Average Score of Annotator 64: 0.35566044647923256
Average Score of Annotator 60: 0.412180736139301
Average Score of Annotator 61: 0.37110000577589375
Average Score of Annotator 63: 0.2818925156965555
```

As we can see above average score of each annotator is >0.2 so I haven't eliminated any annotator decision

In [34]:
```
mode_nyt_stance_0 = df_nyt_stance_0.mode(axis = 1)

df_mode_nyt_stance_0 = pd.DataFrame(data = mode_nyt_stance_0)
df_mode_nyt_stance_0
```

Out[34]:

	0	1
0	unclear	NaN
1	unclear	NaN
2	unclear	NaN
3	unclear	NaN
4	anti-mitigation	pro-mitigation
...	...	...
208	pro-mitigation	NaN
209	unclear	NaN
210	pro-mitigation	NaN
211	anti-mitigation	NaN
212	unclear	NaN

213 rows × 2 columns

Now the reason for having 2 columns is that we might have equal values of "unclear","anti-mitigation" and "pro-mitigation".

Now as defined in the problem we will take HIGHEST AVERAGE of Kappa Score of Annotator for such rows which have equal values.

We have the annotator_60's value as highest which is 0.412180736139301.

In [36]:

```
mode_df = [df_nyt_stance_0, df_mode_nyt_stance_0]
new_mode_df = pd.concat(mode_df, axis = 1)

new_mode_df = new_mode_df.rename(columns = {0:"FinalValue",1:"Value_2"})
new_mode_df
```

Out[36]:

	text	annotation_64	annotation_60	annotation_61	annotation_63	FinalValue	Va
0	To all those who gloat “elections have consequ...	anti-mitigation	unclear	pro-mitigation	unclear	unclear	
1	Corporations and health privacy in the US? How...	unclear	unclear	unclear	anti-mitigation	unclear	
2	No. Done with it.	unclear	unclear	unclear	unclear	unclear	
3	I have not eaten outside my house since March....	unclear	unclear	unclear	pro-mitigation	unclear	
4	More people die each day in US from the virus ...	anti-mitigation	pro-mitigation	pro-mitigation	anti-mitigation	anti-mitigation	miti
...	...	...	...	...	...	...	
208	Require them. Period.\n\nAll the good and dece...	unclear	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	
209	Unvaccinated worker(s) at a Kentucky nursing h...	unclear	unclear	anti-mitigation	pro-mitigation	unclear	
210	The vaccine mandate is about an individual not...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	
211	Fully vaccinated American here. Many of the ex...	unclear	anti-mitigation	anti-mitigation	anti-mitigation	anti-mitigation	
212	It's maddening that certain scientists and hea...	pro-mitigation	unclear	unclear	anti-mitigation	unclear	

213 rows × 7 columns

In [43]:

```
# Setting annotator_60 value to FinalValue column:
for idx,rows in new_mode_df.iterrows():
    if new_mode_df.Value_2[idx] == "unclear" or new_mode_df.Value_2[idx] == "
        new_mode_df.FinalValue[idx] = new_mode_df.annotation_60[idx]
```

In [44]:
```
nyt_stance_0_final = new_mode_df.drop(['Value_2'], axis = 1)
nyt_stance_0_final
```

Out[44]:

	text	annotation_64	annotation_60	annotation_61	annotation_63	FinalValue
0	To all those who gloat “elections have consequ...	anti-mitigation	unclear	pro-mitigation	unclear	unclear
1	Corporations and health privacy in the US? How...	unclear	unclear	unclear	anti-mitigation	unclear
2	No. Done with it.	unclear	unclear	unclear	unclear	unclear
3	I have not eaten outside my house since March....	unclear	unclear	unclear	pro-mitigation	unclear
4	More people die each day in US from the virus ...	anti-mitigation	pro-mitigation	pro-mitigation	anti-mitigation	pro-mitigation
...	...	...	...	...	...	...
208	Require them. Period.\n\nAll the good and dece...	unclear	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
209	Unvaccinated worker(s) at a Kentucky nursing h...	unclear	unclear	anti-mitigation	pro-mitigation	unclear
210	The vaccine mandate is about an individual not...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
211	Fully vaccinated American here. Many of the ex...	unclear	anti-mitigation	anti-mitigation	anti-mitigation	anti-mitigation
212	It's maddening that certain scientists and hea...	pro-mitigation	unclear	unclear	anti-mitigation	unclear

213 rows × 6 columns

2) nyt_stance_1

In [53]:
```
df_nyt_stance_1 = pd.read_csv("C:\\College\\SEM 2\\NLP\\Assignment 2\\nyt_sta
df_nyt_stance_1.isnull().sum()
```

Out[53]:
```
text            0
annotation_34   0
annotation_5    1
annotation_20   1
annotation_4    0
dtype: int64
```

In [54]:
```
df_nyt_stance_1 = df_nyt_stance_1.fillna(value = "missing")
df_nyt_stance_1.isnull().sum()
```

Out[54]:
```
text            0
annotation_34   0
annotation_5    0
annotation_20   0
annotation_4    0
dtype: int64
```

In [58]:
```
nyt_stance_score_1_34_5 = cohen_kappa_score(df_nyt_stance_1.annotation_34, df
nyt_stance_score_1_34_20 = cohen_kappa_score(df_nyt_stance_1.annotation_34,df
nyt_stance_score_1_34_4 = cohen_kappa_score(df_nyt_stance_1.annotation_34,df_
nyt_stance_score_1_5_20 = cohen_kappa_score(df_nyt_stance_1.annotation_5,df_n
nyt_stance_score_1_5_4 = cohen_kappa_score(df_nyt_stance_1.annotation_5,df_ny
nyt_stance_score_1_20_4 = cohen_kappa_score(df_nyt_stance_1.annotation_20,df_

avg_nyt_stance_score_1_34 = (nyt_stance_score_1_34_5 + nyt_stance_score_1_34_
avg_nyt_stance_score_1_5 = (nyt_stance_score_1_34_5 + nyt_stance_score_1_5_20
avg_nyt_stance_score_1_20 = (nyt_stance_score_1_5_20 + nyt_stance_score_1_34_
avg_nyt_stance_score_1_4 = (nyt_stance_score_1_34_4 + nyt_stance_score_1_5_4

print("Average Score of Each Annotator nyt_stance_1 File \n")
print("Average Score of Annotator 34:",avg_nyt_stance_score_1_34)
print("Average Score of Annotator 5:",avg_nyt_stance_score_1_5)
print("Average Score of Annotator 20:",avg_nyt_stance_score_1_20)
print("Average Score of Annotator 4:",avg_nyt_stance_score_1_4)
```

```
Average Score of Each Annotator nyt_stance_1 File

Average Score of Annotator 34: 0.38231525380480935
Average Score of Annotator 5: 0.29590450222106796
Average Score of Annotator 20: 0.34262175463818884
Average Score of Annotator 4: 0.41087478217237505
```

As we can see above average score of each annotator is >0.2 so I haven't eliminated any annotator decision

In [59]:

```
mode_nyt_stance_1 = df_nyt_stance_1.mode(axis = 1)

df_mode_nyt_stance_1 = pd.DataFrame(data = mode_nyt_stance_1)
df_mode_nyt_stance_1
```

Out[59]:

	0	1	2	3	4
0	anti-mitigation	NaN	NaN	NaN	NaN
1	pro-mitigation	NaN	NaN	NaN	NaN
2	unclear	NaN	NaN	NaN	NaN
3	unclear	NaN	NaN	NaN	NaN
4	pro-mitigation	NaN	NaN	NaN	NaN
...	...	...	...	...	...
295	pro-mitigation	NaN	NaN	NaN	NaN
296	pro-mitigation	NaN	NaN	NaN	NaN
297	pro-mitigation	NaN	NaN	NaN	NaN
298	pro-mitigation	NaN	NaN	NaN	NaN
299	pro-mitigation	NaN	NaN	NaN	NaN

300 rows × 5 columns

In [62]:

```
mode_df_1 = [df_nyt_stance_1, df_mode_nyt_stance_1]
new_mode_df_1 = pd.concat(mode_df_1, axis = 1)

new_mode_df_1 = new_mode_df_1.rename(columns = {0:"FinalValue",1:"Value_1",2:
new_mode_df_1
```

Out[62]:

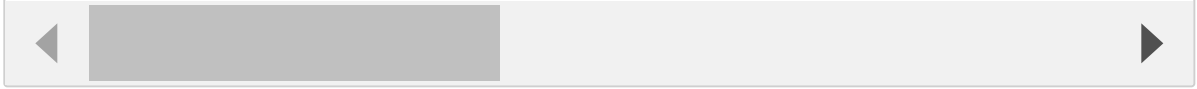
	text	annotation_34	annotation_5	annotation_20	annotation_4	FinalValue	Value_1
0	I'm so tired of the fear mongering. Every vari...	anti-mitigation	anti-mitigation	unclear	pro-mitigation	anti-mitigation	NaN
1	We need a vaccine soonest. Not sure why Russi...	pro-mitigation	unclear	pro-mitigation	pro-mitigation	pro-mitigation	NaN
2	The lack of leadership on this is astonishing	unclear	unclear	unclear	unclear	unclear	NaN
3	Alternative reality is really gaining traction...	pro-mitigation	unclear	unclear	unclear	unclear	NaN
4	I would be much more likely to patronize a bus...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	NaN
...	...	...	...	...	...	...	...
295	The longer we drag this on, the more likely we...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	NaN
296	Interesting that CEOs are worried about the pr...	pro-mitigation	pro-mitigation	anti-mitigation	pro-mitigation	pro-mitigation	NaN
297	I'm rapidly arriving at the opinion that we ne...	pro-mitigation	pro-mitigation	anti-mitigation	pro-mitigation	pro-mitigation	NaN
298	The world we live in is different and riskier...	pro-mitigation	pro-mitigation	unclear	pro-mitigation	pro-mitigation	NaN

	text	annotation_34	annotation_5	annotation_20	annotation_4	FinalValue	Value_1
299	Get a vaccine or get a pink slip. \nThere is ...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	NaN

300 rows × 10 columns

In [65]:

```
# Setting annotator_4 value to FinalValue column:
for idx,rows in new_mode_df_1.iterrows():
    if new_mode_df_1.Value_1[idx] == "unclear" or new_mode_df_1.Value_1[idx]
        new_mode_df_1.FinalValue[idx] = new_mode_df_1.annotation_4[idx]
```



Here I have checked the value of very first column after FinalValue column because our final aim is to replace the conflicting column with highest annotator column. So we can compare any column here and replace it with Highest annotator. Here highest annotator value is of annotator_4.

In [78]:
```
nyt_stance_1_final = new_mode_df_1.drop(['Value_1','Value_2','Value_3','Value
nyt_stance_1_final
```

Out[78]:

	text	annotation_34	annotation_5	annotation_20	annotation_4	FinalValue
0	I’m so tired of the fear mongering. Every vari...	anti-mitigation	anti-mitigation	unclear	pro-mitigation	anti-mitigation
1	We need a vaccine soonest. Not sure why Russi...	pro-mitigation	unclear	pro-mitigation	pro-mitigation	pro-mitigation
2	The lack of leadership on this is astonishing	unclear	unclear	unclear	unclear	unclear
3	Alternative reality is really gaining traction...	pro-mitigation	unclear	unclear	unclear	unclear
4	I would be much more likely to patronize a bus...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
...	...	...	...	...	...	...
295	The longer we drag this on, the more likely we...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
296	Interesting that CEOs are worried about the pr...	pro-mitigation	pro-mitigation	anti-mitigation	pro-mitigation	pro-mitigation
297	I'm rapidly arriving at the opinion that we ne...	pro-mitigation	pro-mitigation	anti-mitigation	pro-mitigation	pro-mitigation
298	The world we live in is different and riskier...	pro-mitigation	pro-mitigation	unclear	pro-mitigation	pro-mitigation
299	Get a vaccine or get a pink slip. \nThere is ...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation

300 rows × 6 columns

3) nyt_stance_2

In [81]:
```
df_nyt_stance_2 = pd.read_csv("C:\\College\\SEM 2\\NLP\\Assignment 2\\nyt_sta
df_nyt_stance_2.isnull().sum()
```

Out[81]:
```
text              0
annotation_7      2
annotation_8    298
annotation_9      1
annotation_78     0
annotation_62     0
dtype: int64
```

In [83]:
```
df_nyt_stance_2 = df_nyt_stance_2.fillna(value = "missing")
df_nyt_stance_2.isnull().sum()
```

Out[83]:
```
text            0
annotation_7    0
annotation_8    0
annotation_9    0
annotation_78   0
annotation_62   0
dtype: int64
```

In [85]:

```
nyt_stance_score_2_7_8 = cohen_kappa_score(df_nyt_stance_2.annotation_7, df_n
nyt_stance_score_2_7_9 = cohen_kappa_score(df_nyt_stance_2.annotation_7,df_ny
nyt_stance_score_2_7_78 = cohen_kappa_score(df_nyt_stance_2.annotation_7,df_n
nyt_stance_score_2_7_62 = cohen_kappa_score(df_nyt_stance_2.annotation_7,df_n
nyt_stance_score_2_8_9 = cohen_kappa_score(df_nyt_stance_2.annotation_8,df_ny
nyt_stance_score_2_8_78 = cohen_kappa_score(df_nyt_stance_2.annotation_8,df_n
nyt_stance_score_2_8_62 = cohen_kappa_score(df_nyt_stance_2.annotation_8,df_n
nyt_stance_score_2_9_78 = cohen_kappa_score(df_nyt_stance_2.annotation_9,df_n
nyt_stance_score_2_9_62 = cohen_kappa_score(df_nyt_stance_2.annotation_9,df_n
nyt_stance_score_2_78_62 = cohen_kappa_score(df_nyt_stance_2.annotation_78,df

avg_nyt_stance_score_2_7 = (nyt_stance_score_2_7_8 + nyt_stance_score_2_7_9 +
avg_nyt_stance_score_2_8 = (nyt_stance_score_2_7_8 + nyt_stance_score_2_8_9 +
avg_nyt_stance_score_2_9 = (nyt_stance_score_2_7_9 + nyt_stance_score_2_8_9 +
avg_nyt_stance_score_2_78 = (nyt_stance_score_2_7_78 + nyt_stance_score_2_8_7
avg_nyt_stance_score_2_62 = (nyt_stance_score_2_7_62 + nyt_stance_score_2_8_6

print("Average Score of Each Annotator nyt_stance_2 File \n")
print("Average Score of Annotator 7:",avg_nyt_stance_score_2_7)
print("Average Score of Annotator 8:",avg_nyt_stance_score_2_8)
print("Average Score of Annotator 9:",avg_nyt_stance_score_2_9)
print("Average Score of Annotator 78:",avg_nyt_stance_score_2_78)
print("Average Score of Annotator 62:",avg_nyt_stance_score_2_62)
```

```
Average Score of Each Annotator nyt_stance_2 File

Average Score of Annotator 7: 0.3089403539517842
Average Score of Annotator 8: -0.0005676090538629663
Average Score of Annotator 9: 0.18790021508095137
Average Score of Annotator 78: 0.31627575307191835
Average Score of Annotator 62: 0.30450723707881616
```

Here we can see that annotator_8 and annotator_9 has value <0.2 so we will drop those annotators

In [88]:
```
df_nyt_stance_2 = df_nyt_stance_2.drop(['annotation_8','annotation_9'], axis
df_nyt_stance_2
```

Out[88]:

	text	annotation_7	annotation_78	annotation_62
0	What do you mean that the pandemic has stretch...	unclear	unclear	unclear
1	I'm not sure if it qualifies as Big Brother wh...	pro-mitigation	unclear	pro-mitigation
2	What if smallpox and polio came roaring back o...	pro-mitigation	pro-mitigation	pro-mitigation
3	I'm continuing to wear a mask. Even in bed, s...	pro-mitigation	pro-mitigation	unclear
4	Although my husband and I really, really miss ...	pro-mitigation	pro-mitigation	unclear
...	...	...	...	...
295	I went out for lunch yesterday with a friend. ...	unclear	pro-mitigation	pro-mitigation
296	Even a year and a half later, I don't understa...	pro-mitigation	pro-mitigation	pro-mitigation
297	Let me get this straight. Vaccinated people ar...	pro-mitigation	unclear	anti-mitigation
298	I have not suffered from a cold or the flu for...	pro-mitigation	pro-mitigation	pro-mitigation
299	I cannot understand why people are still stand...	pro-mitigation	pro-mitigation	unclear

300 rows × 4 columns

In [89]:
```
mode_nyt_stance_2 = df_nyt_stance_2.mode(axis = 1)

df_mode_nyt_stance_2 = pd.DataFrame(data = mode_nyt_stance_2)
df_mode_nyt_stance_2
```

Out[89]:

	0	1	2	3
0	unclear	NaN	NaN	NaN
1	pro-mitigation	NaN	NaN	NaN
2	pro-mitigation	NaN	NaN	NaN
3	pro-mitigation	NaN	NaN	NaN
4	pro-mitigation	NaN	NaN	NaN
...	...	...	...	...
295	pro-mitigation	NaN	NaN	NaN
296	pro-mitigation	NaN	NaN	NaN
297	Let me get this straight. Vaccinated people ar...	anti-mitigation	pro-mitigation	unclear
298	pro-mitigation	NaN	NaN	NaN
299	pro-mitigation	NaN	NaN	NaN

300 rows × 4 columns

In [96]:
```
mode_df_2 = [df_nyt_stance_2, df_mode_nyt_stance_2]
new_mode_df_2 = pd.concat(mode_df_2, axis = 1)

new_mode_df_2 = new_mode_df_2.rename(columns = {0:"FinalValue",1:"Value_1",2:
new_mode_df_2
```

Out[96]:

	text	annotation_7	annotation_78	annotation_62	FinalValue	Value_1	Value_2	V
0	What do you mean that the pandemic has stretch...	unclear	unclear	unclear	unclear	NaN	NaN	
1	I'm not sure if it qualifies as Big Brother wh...	pro-mitigation	unclear	pro-mitigation	pro-mitigation	NaN	NaN	
2	What if smallpox and polio came roaring back o...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	NaN	NaN	
3	I'm continuing to wear a mask. Even in bed, s...	pro-mitigation	pro-mitigation	unclear	pro-mitigation	NaN	NaN	
4	Although my husband and I really, really miss ...	pro-mitigation	pro-mitigation	unclear	pro-mitigation	NaN	NaN	
...	...	...	...	...	...	...	...	
295	I went out for lunch yesterday with a friend. ...	unclear	pro-mitigation	pro-mitigation	pro-mitigation	NaN	NaN	
296	Even a year and a half later, I don't understa...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	NaN	NaN	
297	Let me get this straight. Vaccinated people ar...	pro-mitigation	unclear	anti-mitigation	Let me get this straight. Vaccinated people ar...	anti-mitigation	pro-mitigation	

	text	annotation_7	annotation_78	annotation_62	FinalValue	Value_1	Value_2	\
298	I have not suffered from a cold or the flu for...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation	NaN	NaN	
299	I cannot understand why people are still stand...	pro-mitigation	pro-mitigation	unclear	pro-mitigation	NaN	NaN	

300 rows × 8 columns

In [104]:

```
# Setting annotator_78 value to FinalValue column:
for idx,rows in new_mode_df_2.iterrows():
    if new_mode_df_2.Value_1[idx] == "unclear" or new_mode_df_2.Value_1[idx]
        new_mode_df_2.FinalValue[idx] = new_mode_df_2.annotation_78[idx]
```

In [105]:
```
nyt_stance_2_final = new_mode_df_2.drop(['Value_1', 'Value_2', 'Value_3'], ax
nyt_stance_2_final
```

Out[105]:

	text	annotation_7	annotation_78	annotation_62	FinalValue
0	What do you mean that the pandemic has stretch...	unclear	unclear	unclear	unclear
1	I'm not sure if it qualifies as Big Brother wh...	pro-mitigation	unclear	pro-mitigation	pro-mitigation
2	What if smallpox and polio came roaring back o...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
3	I'm continuing to wear a mask. Even in bed, s...	pro-mitigation	pro-mitigation	unclear	pro-mitigation
4	Although my husband and I really, really miss ...	pro-mitigation	pro-mitigation	unclear	pro-mitigation
...	...	...	...	...	...
295	I went out for lunch yesterday with a friend. ...	unclear	pro-mitigation	pro-mitigation	pro-mitigation
296	Even a year and a half later, I don't understa...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
297	Let me get this straight. Vaccinated people ar...	pro-mitigation	unclear	anti-mitigation	unclear
298	I have not suffered from a cold or the flu for...	pro-mitigation	pro-mitigation	pro-mitigation	pro-mitigation
299	I cannot understand why people are still stand...	pro-mitigation	pro-mitigation	unclear	pro-mitigation

300 rows × 5 columns

Now that we have all the files ready we will merge Final Dataframe of nyt_stance_0, nyt_stance_1 & nyt_stance_2

In [106]:
```
# Merging all the files
final_df_primary = pd.concat([nyt_stance_0_final,nyt_stance_1_final,nyt_stanc
keep_columns = ['text','FinalValue']
final_df_primary = final_df_primary[keep_columns]
final_df_primary = final_df_primary.rename(columns = {"FinalValue":"label"})
final_df_primary.to_csv(r"C:\College\SEM 2\NLP\Assignment 2\Final Dataset\Fin
```