In [1]:

```
import pandas as pd
```

In [2]:

```
data = pd.read_csv('fulltext_cleaned.csv')
```

In [3]:

```
data.columns
```

Out[3]:

```
Index(['uid', 'full_text', '5g', 'abortion', 'aliens', 'antiasian',
       'antiblack', 'antiimmigrant', 'antilatinx', 'antilgbt', 'antimuslim',
       'antisemitic', 'antivaxx', 'biden', 'bigtech', 'climatedenial',
       'coronavirus', 'criticalracetheory', 'misogyny', 'presidentbiden',
       'pseudoscience', 'qanon', 'voterfraud', 'votinglaws', 'whitesupremacy'],
      dtype='object')
```

In [6]:

```
compressed = data.iloc[:,2:].idxmax(axis=1)
```

In [14]:

```
data_compressed = pd.DataFrame(compressed, columns=['label'])
```

In [15]:

```
data_compressed
```

Out[15]:

	label
0	voterfraud
1	whitesupremacy
2	antilatinx
3	biden
4	bigtech
...	...
108920	coronavirus
108921	presidentbiden
108922	antiasian
108923	biden
108924	voterfraud

108925 rows × 1 columns

In [17]:

```
df = data.iloc[:,:2].join(data_compressed)
```

In [33]:

```
df
```

Out[33]:

	uid	full_text	label

	uid	full_text	label
0	380117	michigan secretary state jocelyn benson pictur...	voterfraud
1	448712	joe hall former marine handyman currently lay ...	whitesupremacy
2	256646	politics entertainment media economy world lon...	antilatinx
3	256646	politics entertainment media economy world lon...	biden
4	406930	thomas consider conservative high court make p...	bigtech
...	...	...	...
108920	402163	chinese communist party s cyberspace administr...	coronavirus
108921	402163	chinese communist party s cyberspace administr...	presidentbiden
108922	402163	chinese communist party s cyberspace administr...	antiasian
108923	319881	president joe biden cancel monday trip state d...	biden
108924	188013	days reveal 2020 us elections rig canadian cro...	voterfraud

108925 rows × 3 columns

In [20]:

```
label_total_count = pd.DataFrame(df.groupby('label').size())
```

In [23]:

```
label_total_count.reset_index(inplace=True)
```

In [31]:

```
label_total_count.rename(columns = {0:'total_label_count'}, inplace=True)
```

In [32]:

```
label_total_count
```

Out[32]:

	label	total_label_count
0	5g	243
1	abortion	1519
2	aliens	14
3	antiasian	789
4	antiblack	1128
5	antiimmigrant	3687
6	antilatinx	4222
7	antilgbt	4006
8	antimuslim	2744
9	antisemitic	1457
10	antivaxx	304
11	biden	40165
12	bigtech	3986
13	climatedenial	156
14	coronavirus	12440
15	criticalracetheory	28
16	misogyny	656
17	presidentbiden	2631
18	pseudoscience	341

	label	total_label_count
19	qanon	271
20	voterfraud	19093
21	votinglaws	120
22	whitesupremacy	8925

In [34]:

```
from sklearn.model_selection import train_test_split
train, test = train_test_split(df, test_size=0.2,stratify=df['label'])
```

In [37]:

```
train_label_count = pd.DataFrame(train.groupby('label').size())
train_label_count.reset_index(inplace=True)
```

In [39]:

```
train_label_count.rename(columns={0:'train_count'}, inplace=True)
```

In [40]:

```
test_label_count = pd.DataFrame(test.groupby('label').size())
test_label_count.reset_index(inplace=True)
```

In [41]:

```
test_label_count.rename(columns={0:'test_count'}, inplace=True)
```

In [51]:

```
label_count = label_total_count.merge(train_label_count.merge(test_label_count)).reset_i
ndex()
```

In [53]:

```
label_count.rename(columns={'index':'label_index'}, inplace=True)
```

In [55]:

```
label_count['label_index'] = label_count['label_index'] + 1
```

In [56]:

```
label_count
```

Out[56]:

	label_index	label	total_label_count	train_count	test_count
0	1	5g	243	194	49
1	2	abortion	1519	1215	304
2	3	aliens	14	11	3
3	4	antiasian	789	631	158
4	5	antiblack	1128	902	226
5	6	antiimmigrant	3687	2950	737
6	7	antilatinx	4222	3378	844
7	8	antilgbt	4006	3205	801
8	9	antimuslim	2744	2195	549
9	10	antisemitic	1457	1166	291
10	11	antivaxx	304	243	61
11	12	biden	40165	32132	8033

	label_index	label	total_label_count	train_count	test_count
12	13	bigtech	3986	3189	797
13	14	climatedenial	156	125	31
14	15	coronavirus	12440	9952	2488
15	16	criticalracetheory	28	22	6
16	17	misogyny	656	525	131
17	18	presidentbiden	2631	2105	526
18	19	pseudoscience	341	273	68
19	20	qanon	271	217	54
20	21	voterfraud	19093	15274	3819
21	22	votinglaws	120	96	24
22	23	whitesupremacy	8925	7140	1785

In [57]:

```
label_count.to_csv('label_count.csv', index=False)
```

In [59]:

```
train.reset_index(inplace=True, drop=True)
```

In [61]:

```
test.reset_index(inplace=True, drop=True)
```

In [62]:

```
train_one_hot = pd.get_dummies(train['label'])
train.drop(columns = 'label', axis=1, inplace=True)
train_set = train.join(train_one_hot)
```

```
/opt/anaconda3/lib/python3.8/site-packages/pandas/core/frame.py:4906: SettingWithCopyWarn
ing:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_g
uide/indexing.html#returning-a-view-versus-a-copy
  return super().drop(
```

In [63]:

```
train_set
```

Out[63]:

	uid	full_text	5g	abortion	aliens	antiasian	antiblack	antiimmigrant	antilatinx	antilgbt	...	climatedenial	c
0	445020	long post key insight greatly inspire encourag...	0	0	0	0	0	0	0	0	...	0	
1	316622	arizona state senate announce friday hire inde...	0	0	0	0	0	0	0	0	...	0	
2	225503	washington post thursday promote create politi...	0	0	0	0	0	0	0	0	...	0	
3	253398	politics entertainment media	0	0	0	0	0	0	0	0	...	0	

	uid	full_text	5g	abortion	aliens	antiasian	antiblack	antiimmigrant	antilatinx	antilgbt	...	climatedenial	c
		economy world lon...											
4	276629	biden soon president jihad resume iran parliam...	0	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
87135	344961	posters mocking biden fa uci covid mandates ap...	0	0	0	0	0	0	0	0	...	0	
87136	171143	jim h of t founder editor gateway pundit one t...	0	0	0	0	0	0	0	0	...	0	
87137	629035	moment every donald trump opponent wait hand j...	0	0	0	0	0	0	0	0	...	0	
87138	245640	oath keepers non-partisan association current ...	0	0	0	0	0	0	0	0	...	0	
87139	147311	new listen fox news article owner delaware com...	0	0	0	0	0	0	0	0	...	0	

87140 rows × 25 columns

In [64]:

```
test_one_hot = pd.get_dummies(test['label'])
test.drop(columns = 'label', axis=1, inplace=True)
test_set = test.join(test_one_hot)
```

```
/opt/anaconda3/lib/python3.8/site-packages/pandas/core/frame.py:4906: SettingWithCopyWarn
ing:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_g
uide/indexing.html#returning-a-view-versus-a-copy
  return super().drop(
```

In [65]:

```
test_set
```

Out[65]:

	uid	full_text	5g	abortion	aliens	antiasian	antiblack	antiimmigrant	antilatinx	antilgbt	...	climatedenial	c
0	305823	president joe biden slash thousand union job f...	0	0	0	0	0	0	0	0	...	0	
1	897854	stacey abrams-founded group sues georgia voter...	0	0	0	0	0	0	0	0	...	0	
		tuesday 3											

	uid	full_text	5g	abortion	aliens	antiasian	antiblack	antiimmigrant	antilatinx	antilgbt	...	climatedenial	co
2	200770	tuesday 5 pm eastern dr shiva ayya dura i joe ...	0	0	0	0	0	0	0	0	...	0	
3	201901	encode utf-8 2016 election republicans publicl...	0	0	0	0	0	0	0	0	...	0	
4	237478	official never hear doctor argument debate soo...	0	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
21780	413909	im white think im racist that s bs eric bollin...	0	0	0	0	0	0	0	0	...	0	
21781	185142	far spontaneous mythical today militant politi...	0	0	0	0	0	0	0	0	...	0	
21782	218608	left wing medium outlet vox publish piece tues...	0	0	0	0	0	0	0	0	...	0	
21783	399665	trump blast joe biden blistering statement wed...	0	0	0	0	0	0	0	0	...	0	
21784	472468	pete iowa three word 2020 fill pride day born ...	0	0	0	0	0	0	0	1	...	0	

21785 rows × 25 columns

In [66]:

```
train_set.to_csv('fulltext_cleaned_train.csv', index=False)
test_set.to_csv('fulltext_cleaned_test.csv', index=False)
```

In [ ]: