In [1]:
```
import pandas as pd
import yaml
import matplotlib.pyplot as plt
from IPython.display import display, HTML
```

In [2]:
```
from modules.data import Data
from modules.search import Search
from modules.video import Video
from modules.analyze import Analyze
```

In [3]:
```
data_obj = Data()
analyze_obj = Analyze()
```

In [4]:
```
df_video_labeled = pd.read_csv("unique_id_map/videos_anonymized.csv", dtype={"ch
```

In [5]:
```
# Display duration in a readable format
df_video_labeled["video_duration"] = df_video_labeled["video_duration"].apply(ar
df_video_labeled.head()
# Get engagement metrics
df_video_labeled["likes_to_dislikes"] = df_video_labeled.apply(lambda row: analy
df_video_labeled["dislikes_to_likes"] = df_video_labeled.apply(lambda row: analy
df_video_labeled["engagement_score"] = df_video_labeled.apply(lambda row: analyz
```

In [6]:
```
dict_variables = data_obj.load_yaml("variables.yaml")
list_category = dict_variables["category"]
list_theme = dict_variables["theme"]
```

In [7]:
```
# Get dataframes per category and label
list_df_category, list_df_theme = analyze_obj.splice_by_labels(df_video_labeled,
display(list_df_category[0].head())
```

	video_title	video_description	view_count	like_count	dislike_count	favorite_count	comment_cou
31	Redwood City School District To Install Vape D...	The Redwood City School District Board of Trus...	1105372	24119	1104	0	41
40	Vaping / E-Cigarette Associated Lung Injury: C...	An important update on E-Cigarette / Vaping pr...	17800	459	10	0	1(
41	Vaping / E-Cigarette Lung Failure, Illness, Di...	Please see our most recent update to vaping as...	147156	1335	422	0	8

	video_title	video_description	view_count	like_count	dislike_count	favorite_count	comment_cou
43	The dangers of vaping CBD oil	Dr. Cass Ingram, author of "The Hemp Oil Mirac...	39012	285	421	0	1
44	Vaping vs. Smoking	What are the effects of smoking in the lungs? ...	471	5	3	0	

Stats on views

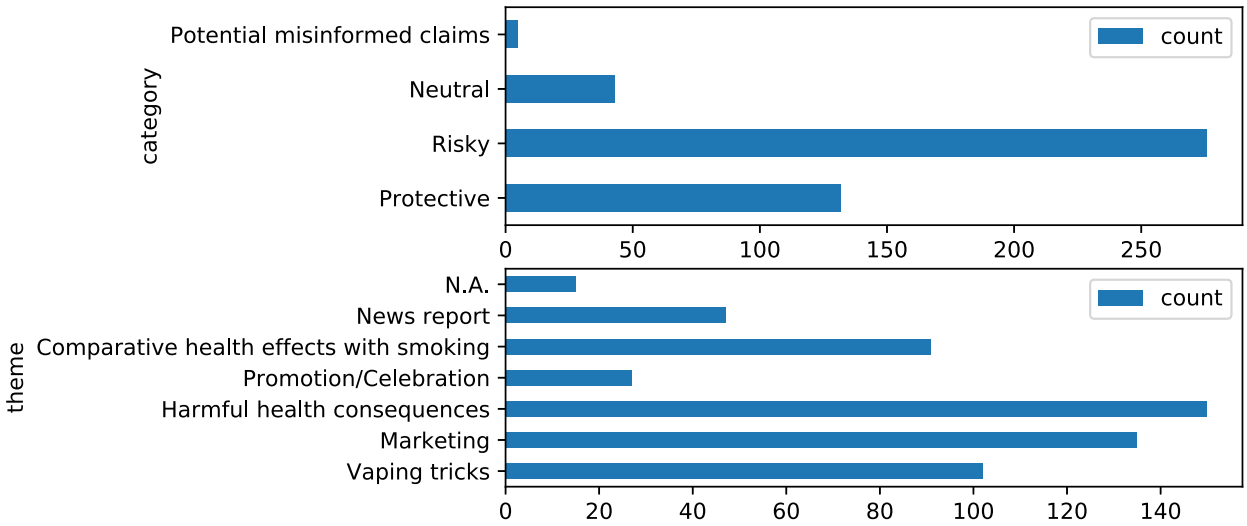
In [8]:

```
df_view_count_category_describe = analyze_obj.describe_df(list_df=list_df_categ
display(df_view_count_category_describe)
df_view_count_theme_describe = analyze_obj.describe_df(list_df=list_df_theme, la
display(df_view_count_theme_describe)

fig, axes = plt.subplots(nrows=2, ncols=1)
df_view_count_category_describe.plot.barh(x="category", y="count", ax=axes[0])
df_view_count_theme_describe.plot.barh(x="theme", y="count", ax=axes[1])
```

	category	count	mean	std	median
0	Protective	132	8.415775e+05	3.478989e+06	46437.5
1	Risky	276	1.492521e+06	4.123480e+06	148152.5
2	Neutral	43	5.698636e+05	1.644923e+06	87154.0
3	Potential misinformed claims	5	1.631372e+05	9.440142e+04	203830.0

	theme	count	mean	std	median
0	Vaping tricks	102	2.972450e+06	6.252307e+06	652430.5
1	Marketing	135	3.542647e+05	7.795443e+05	91488.0
2	Harmful health consequences	150	7.757108e+05	3.277054e+06	47379.0
3	Promotion/Celebration	27	1.965180e+06	2.740716e+06	612535.0
4	Comparative health effects with smoking	91	1.045678e+06	4.185862e+06	33146.0
5	News report	47	6.001753e+05	1.638318e+06	56047.0
6	N.A.	15	3.093649e+05	7.285282e+05	51524.0

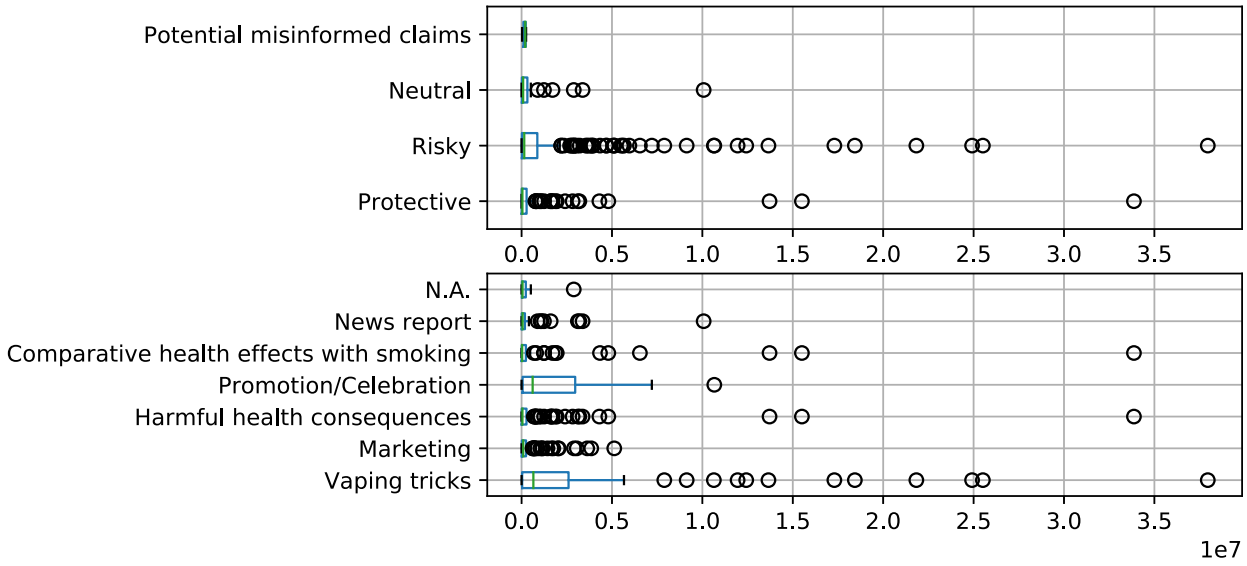
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff6ea242f40>



In [9]:

```
list_view_count_category = [list_df_category[index]["view_count"] for index, _ i
df_view_count_category_boxplot = pd.concat(list_view_count_category, axis=1, key

list_view_count_theme = [list_df_theme[index]["view_count"] for index, _ in enum
df_view_count_theme_boxplot = pd.concat(list_view_count_theme, axis=1, keys=list

fig, axes = plt.subplots(nrows=2, ncols=1)
df_view_count_category_boxplot.boxplot(column=list_category, ax=axes[0], vert=Fa
df_view_count_theme_boxplot.boxplot(column=list_theme, ax=axes[1], vert=False)
```

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff6ea0fd430>



Stats on duration

In [10]:

```
df_video_duration_category_describe = analyze_obj.describe_df(list_df=list_df_ca
display(df_video_duration_category_describe)
df_video_duration_theme_describe = analyze_obj.describe_df(list_df=list_df_theme
display(df_video_duration_theme_describe)
```

	category	mean	std	median

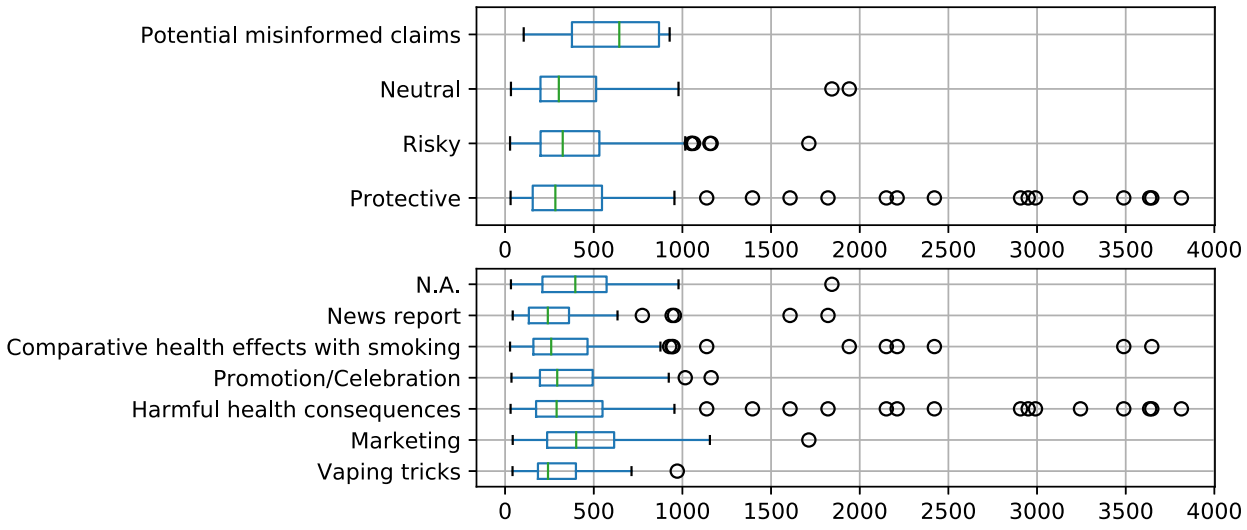
	category	mean	std	median
0	Protective	568.469697	818.485433	283.5
1	Risky	387.420290	256.083333	325.0
2	Neutral	423.232558	407.279057	303.0
3	Potential misinformed claims	584.400000	344.558123	644.0

	theme	mean	std	median
0	Vaping tricks	307.990196	178.662384	242.0
1	Marketing	452.829630	278.239603	401.0
2	Harmful health consequences	552.100000	775.215828	290.5
3	Promotion/Celebration	375.962963	282.919325	294.0
4	Comparative health effects with smoking	481.582418	649.459691	260.0
5	News report	341.553191	361.091028	241.0
6	N.A.	465.600000	459.109511	396.0

In [11]:

```
list_video_duration_category = [list_df_category[index]["video_duration"] for ir
df_video_duration_category = pd.concat(list_video_duration_category, axis=1, key

list_video_duration_theme = [list_df_theme[index]["video_duration"] for index, _
df_video_duration_theme = pd.concat(list_video_duration_theme, axis=1, keys=list

fig, axes = plt.subplots(nrows=2, ncols=1)
df_video_duration_category.boxplot(column=list_category, ax=axes[0], vert=False)
df_video_duration_theme.boxplot(column=list_theme, ax=axes[1], vert=False)
```

Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff6e8779850>



Stats on engagement

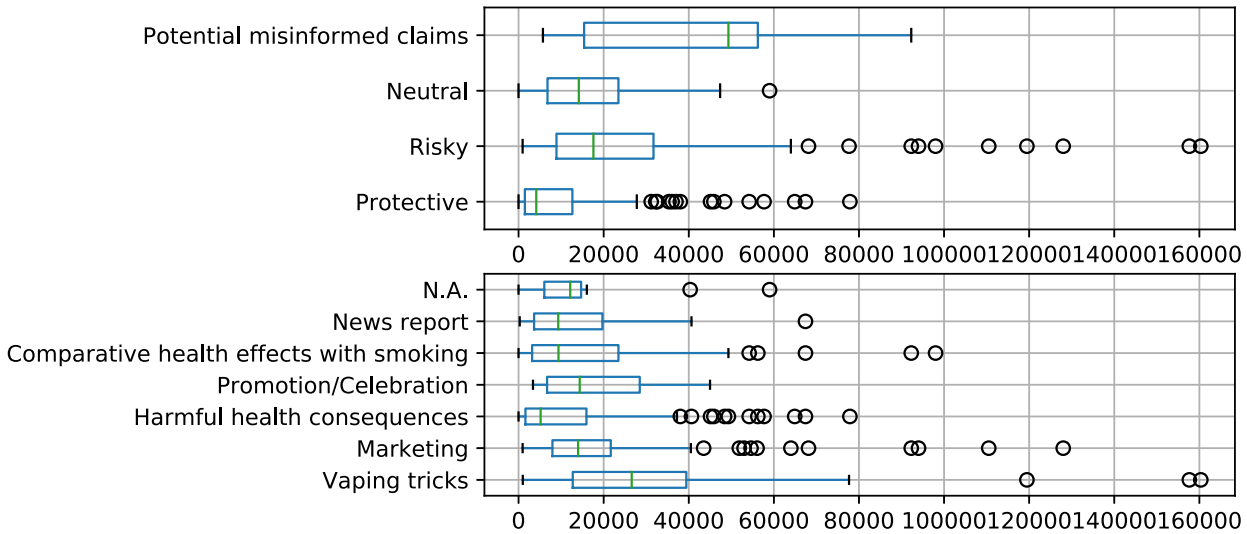
In [12]:

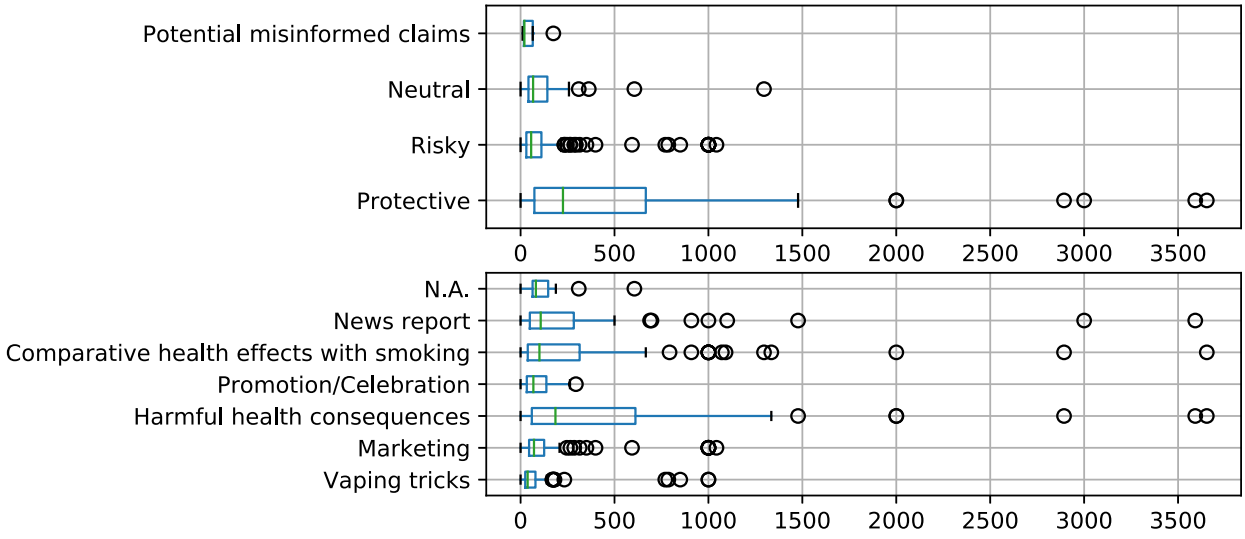
```
# Likes to dislikes
print("Likes to dislikes")
list_likes_to_dislikes_category = [list_df_category[index]["likes_to_dislikes"]
```

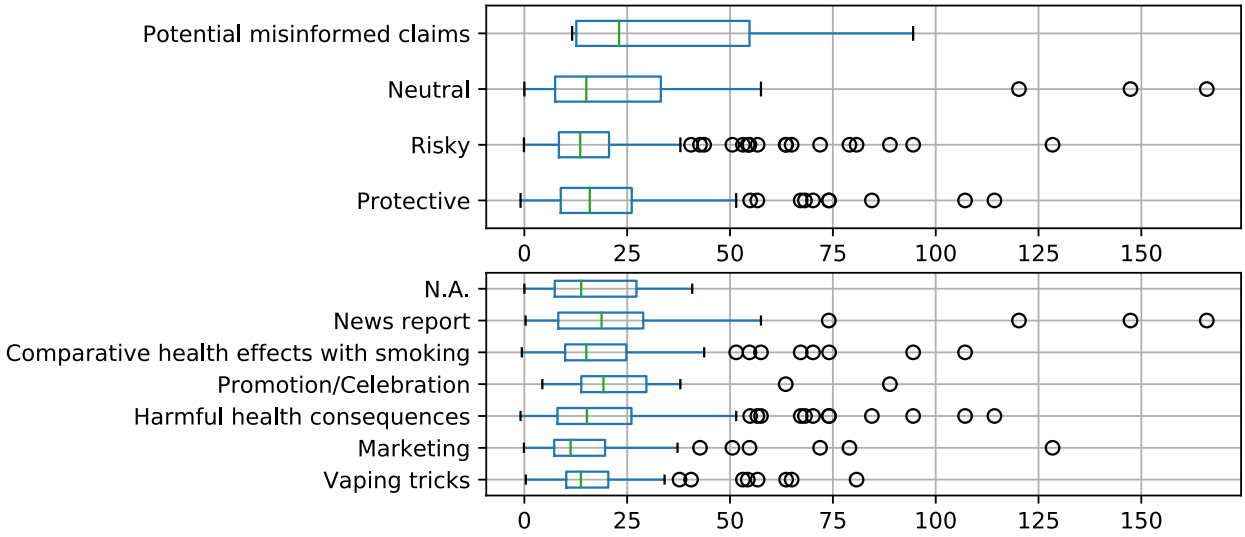
```
df_likes_to_dislikes_category = pd.concat(list_likes_to_dislikes_category, axis=

list_likes_to_dislikes_theme = [list_df_theme[index]["likes_to_dislikes"] for in
df_likes_to_dislikes_theme = pd.concat(list_likes_to_dislikes_theme, axis=1, key

fig, axes = plt.subplots(nrows=2, ncols=1)
df_likes_to_dislikes_category.boxplot(column=list_category, ax=axes[0], vert=Fal
df_likes_to_dislikes_theme.boxplot(column=list_theme, ax=axes[1], vert=False)

# Dislikes to likes
print("Dislikes to likes")

list_dislikes_to_likes_category = [list_df_category[index]["dislikes_to_likes"]
df_dislikes_to_likes_category = pd.concat(list_dislikes_to_likes_category, axis=

list_dislikes_to_likes_theme = [list_df_theme[index]["dislikes_to_likes"] for in
df_dislikes_to_likes_theme = pd.concat(list_dislikes_to_likes_theme, axis=1, key

fig, axes = plt.subplots(nrows=2, ncols=1)
df_dislikes_to_likes_category.boxplot(column=list_category, ax=axes[0], vert=Fal
df_dislikes_to_likes_theme.boxplot(column=list_theme, ax=axes[1], vert=False)

# Engagement score
print("Engagement score")

list_engagement_score_category = [list_df_category[index]["engagement_score"] fo
df_engagement_score_category = pd.concat(list_engagement_score_category, axis=1,

list_engagement_score_theme = [list_df_theme[index]["engagement_score"] for inde
df_engagement_score_theme = pd.concat(list_engagement_score_theme, axis=1, keys=

fig, axes = plt.subplots(nrows=2, ncols=1)
df_engagement_score_category.boxplot(column=list_category, ax=axes[0], vert=Fals
df_engagement_score_theme.boxplot(column=list_theme, ax=axes[1], vert=False)
```

```
Likes to dislikes
Dislikes to likes
Engagement score
```

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff6e83d2e20>




Potential misinformed claims
Neutral
Risky
Protective
0
500
1000
1500
2000
2500
3000
3500
N.A.
News report
Comparative health effects with smoking
Promotion/Celebration
Harmful health consequences
Marketing
Vaping tricks
0
500
1000
1500
2000
2500
3000
3500


Potential misinformed claims
Neutral
Risky
Protective
0
25
50
75
100
125
150
N.A.
News report
Comparative health effects with smoking
Promotion/Celebration
Harmful health consequences
Marketing
Vaping tricks
0
25
50
75
100
125
150