Dataset Analysis

May 13, 2022

[2]:
```
import datasets
import pandas as pd
```

```
/usr/lib/python3/dist-packages/requests/__init__.py:89:
RequestsDependencyWarning: urllib3 (1.26.9) or chardet (3.0.4) doesn't match a
supported version!
  warnings.warn("urllib3 ({}) or chardet ({}) doesn't match a supported "
```

[13]:
```
hsol = datasets.load_dataset("hate_speech_offensive", split="train") \
    .to_pandas() \
    .drop(columns=["count", "hate_speech_count", "offensive_language_count",␣
 ↪"neither_count"])

hsol
```

```
Using custom data configuration default
Reusing dataset hate_speech_offensive (/home/ubuntu/.cache/huggingface/datasets/
hate_speech_offensive/default/1.0.0/5f5dfc7b42b5c650fe30a8c49df90b7dbb9c7a4b3fe4
3ae2e66fabfea35113f5)
```

[13]:

	class	tweet
0	2	!!! RT @mayasolovely: As a woman you shouldn't…
1	1	!!!!! RT @mleew17: boy dats cold…tyga dwn ba…
2	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby…
3	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo…
4	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you…
…	…	…
24778	1	you's a muthaf***in lie “@LifeAsKing: @2…
24779	2	you've gone and broke the wrong heart baby, an…
24780	1	young buck wanna eat!!.. dat nigguh like I ain…
24781	1	youu got wild bitches tellin you lies
24782	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col…

[24783 rows x 2 columns]

[19]:
```
hsol["class"].value_counts() / len(hsol["class"])
```

[19]:
```
1    0.774321
2    0.167978
0    0.057701
Name: class, dtype: float64
```

[38]:
```
for s in hsol[hsol["class"] == 2].sample(n=3, random_state=685).tweet:
    print(s)
```

```
@Pepper_Redbone @Yankees @Mets Oh yeah. And the annoying damn duck calls?? They
outta be banned. Duck horns??
Whoo? A new attraction in Virginia City, birds of prey. Meet this owl famous for
his role in Harry Potter. http://t.co/1NFSh5sCcR
99 percent of the trash we dump in the sea is missing. This is not a good thing
&#8594; http://t.co/CwC2LWCV3T http://t.co/YOkt02HLm1
```

[14]:
```
sarc = pd.read_csv("../SARC2/sarc_processed.csv") \
    .drop(columns=["Unnamed: 0"]) \
    .sample(frac=0.5, random_state=685)

sarc
```

[14]:

	label	text
38950	0	Drop some pistols up there x)
44045	0	Why does he sound like a cross between a Kenne…
48311	0	Once again police are above the laws they are …
56506	0	the first missle shot killed him
50719	1	Yea but its Israel and so it isn't a terror at…
…	…	…
47320	1	NYXL master race
43707	0	Or building more transit options
6955	1	Creepy legal rape is so much better than just …
32137	0	Don't backpedal :-\^)
50890	1	So unthankful, what about the Unidentified Dye?

```
[32333 rows x 2 columns]
```

[21]:
```
sarc["label"].value_counts() / len(sarc["label"])
```

[21]:
```
0    0.503294
1    0.496706
Name: label, dtype: float64
```

[30]:
```
for s in sarc[sarc["label"] == 0].sample(n=3, random_state=686).text:
    print(s)
```

```
Are they making an announcement tomorrow tuesday?
Castle Storm is a really good game, which I know because I bought on the PC long
```

```
ago
Boy oh *boy* those coming debates with Trump will be fun.
```

[28]:
```
for s in sarc[sarc["label"] == 1].sample(n=3, random_state=685).text:
    print(s)
```

```
So the people tasked with stopping gangs and gang violence are nothing but the
very thing they are tasked with stopping, my heart be still!
Thank god fb will tell me whats real and what not
Only if it comes with a Lamborghini, then yes, yes sir you are.
```

[5]:
```
impl = pd.read_table('../implicit_hate_v1_stg2_posts.tsv') \
    .rename(columns={"post": "text", "implicit_class": "label"}) \
    .drop("extra_implicit_class", axis=1) \

impl
```

[5]:

	text	label
0	" : jewish harvard professor noel ignatiev w…	white_grievance
1	rt : tune in today's jmt for my interview w /…	irony
2	: : hate posters ? you want hate buddy ? read …	white_grievance
3	: hunt for white women : black racist serial k…	stereotypical
4	: looks like he was a fan of the islamic jew-h…	stereotypical
…	…	…
6341	Seeing prostitutes is morally wrong, but being…	irony
6342	I wonder how many females I raped today I spen…	irony
6343	Having a criminal record is more attractive to…	stereotypical
6344	Another advantage of being a female: getting b…	inferiority
6345	If men are so privileged and women have it so …	irony

[6346 rows x 2 columns]

[6]:
```
impl["label"].value_counts() / len(impl["label"])
```

[6]:
```
white_grievance    0.242357
incitement         0.199968
stereotypical      0.178538
inferiority        0.135991
irony              0.125591
threatening        0.104948
other              0.012606
Name: label, dtype: float64
```

[27]:
```
for s in impl[impl["label"] == "other"].sample(n=1, random_state=684).text:
    print(s)
```

```
white nationalism is a religion of peace .
```