data_augmentation

April 26, 2023

```
[ ]: import os
     import pandas as pd

     folder_path = '/content/drive/MyDrive/textanalytics/Project/LOLLabeled'
     labeled_path = "/content/drive/MyDrive/textanalytics/Project/emote-controlled/
      ↪data/labeled_dataset.csv"
     col = ["sentiment", "message"]
     df = pd.read_csv(labeled_path)
     df.drop(columns=["date", "channel", "game", "user", "mod", "subscriber"],␣
      ↪inplace=True)
```

```
[ ]: df.rename(columns = {'message':'comment'}, inplace = True)
     df = df[df['sentiment'] != "sentiment"]
     df = df.reset_index(drop=True)

     df = df[df['comment'] != "comment"]
     df = df.reset_index(drop=True)
     df = df.dropna()
     df.comment = df.comment.str.lower()
```

```
[ ]: # only those streamers were taken to prune their dataset because where the ones␣
      ↪with the biggest dataset
     l = []
     l.append("/content/drive/MyDrive/textanalytics/Project/LOLLabeled/
      ↪C9Sneaky_comments.csv")
     l.append("/content/drive/MyDrive/textanalytics/Project/LOLLabeled/
      ↪Yassuo_comments.csv")
     l.append("/content/drive/MyDrive/textanalytics/Project/LOLLabeled/
      ↪TFBlade_comments.csv")
     l.append("/content/drive/MyDrive/textanalytics/Project/LOLLabeled/
      ↪lolTyler1_comments.csv")
```

```
[ ]: for d in l:
       with open(d, "r") as file:
           lines = file.readlines()

       new_lines = []
```

```
  for line in lines:
      new_line = line.replace(";", "")
      new_lines.append(new_line)

  with open(d, "w") as file:
    file.writelines(new_lines)
```

[ ]:
```
col = ["sentiment", "comment"]
tot = pd.DataFrame()
tot = pd.concat([tot, df], axis=0, ignore_index=True)

for file in l:
  lol = pd.read_csv(file)

  lol = lol[lol['sentiment'] != "sentiment"]
  lol = lol.reset_index(drop=True)

  lol = lol[lol['comment'] != "comment"]
  lol = lol.reset_index(drop=True)
  lol = lol.dropna()

  print("prima ", len(lol))
  subset = lol.sample(axis=0, frac=0.5)
  tot = pd.concat([tot, subset], axis=0, ignore_index=True)
  lol = lol.drop(subset.index)
  lol.to_csv(file, index=False)

  print("dopo", len(lol))
  print("df adesso: ", len(tot))
```

```
prima  110691
dopo 55345
df adesso:  57268
prima  116598
dopo 58299
df adesso:  115567
prima  191294
dopo 95647
df adesso:  211214
prima  316309
dopo 158155
df adesso:  369368
```

[ ]:
```
p = "/content/drive/MyDrive/textanalytics/Project/dataset/
 ↪labeled_dataset_augmented.csv"
tot.to_csv(p, index=False)
```

```
[ ]: aug = pd.read_csv(p)
```

```
[ ]: aug.head()
```

```
[ ]:    sentiment                                            comment
     0          1                                          omegalul
     1          1  pepel clap pepel clap pepel clap pepel clap pe…
     2          1                                             zulul
     3         -1                                     cata dumbass
     4          0                                                d:
```

```
[ ]: print(len(aug))
```

```
369368
```