گزارش پروژه پایانی درس پردازش زبان طبیعی

موضوع پروژه: تشخیص نظرات نامناسب
اعضای گروه: سبا رایحی - سهراب نمازی نیا
استاد درس: آقای دکتر مینایی
منتور گروه: هادی شیخی

بخش اول تولید جملات :

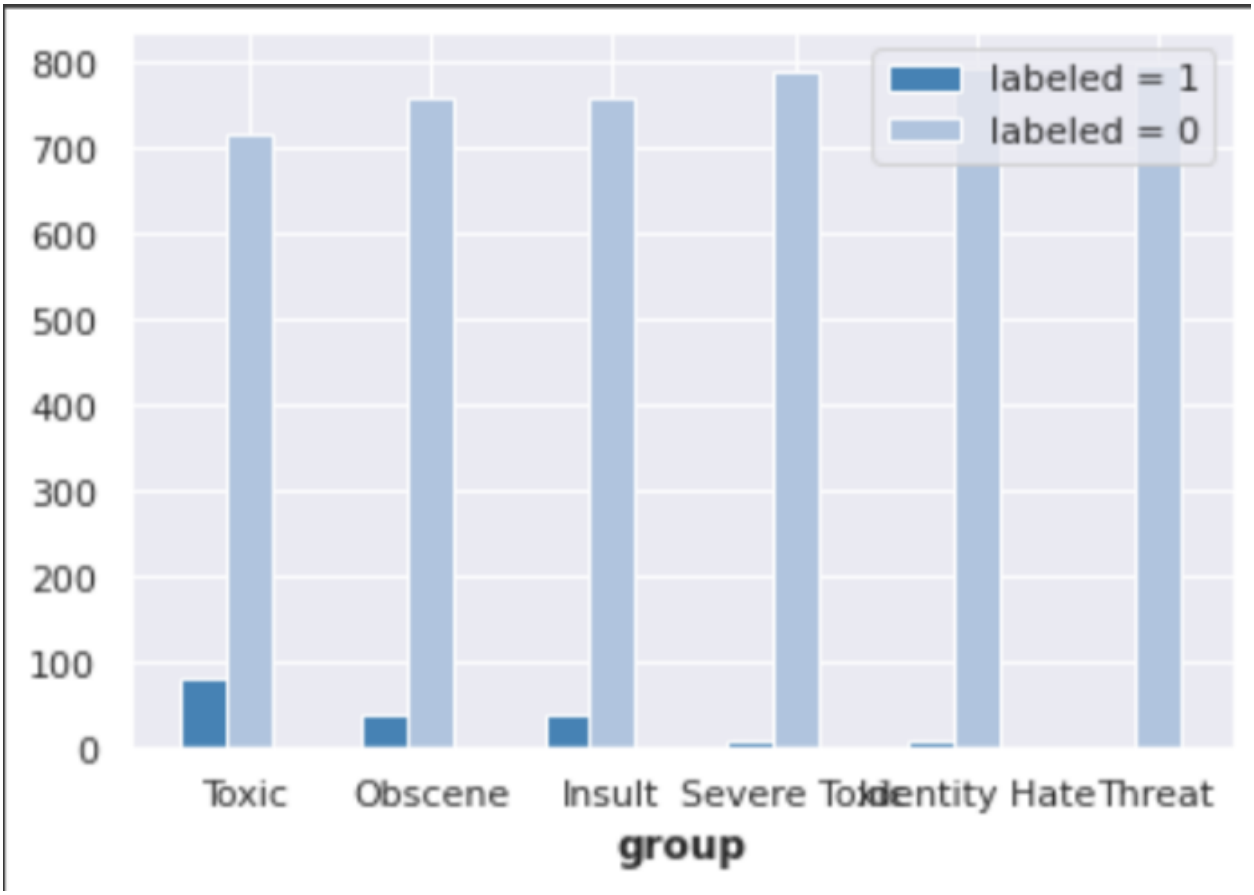
در این بخش ابتدا داده های خود را به دو دسته براساس label تقسیم کرده ایم داده هی ما شامل دو label : برچسب 1 و برچسب 0 است.
برای هر کدام به صورت جداگانه fine tune کردیم برای این کار دیتا را بر اساس برچسب در دو فایل جداگانه می ریزیم که در فایل دیتا آن را قرار داده ایم و بعد dataset را لود کردیم و بعد از آن map به دیتاست را انجام دادیم و از `AutoModelForMaskedLM.from_pretrained`

استفاده کردیم و مدل 'distilroberta-base' را استفاده کردیم و سپس مدل را train کرده و یک تابع برای save مدل مان قرار دادیم که در آن با استفاده از torch. save مدل را با قالب بیان شده در داک save می کنیم و برای هر کدام از برچسب هایمان به طور جداگانه اجرا کردیم

Epoch	Training Loss	Validation Loss
1	13.078600	5.993432
2	6.380400	6.281621
3	7.511600	5.948568
4	6.949400	5.410285
5	4.459000	5.048048

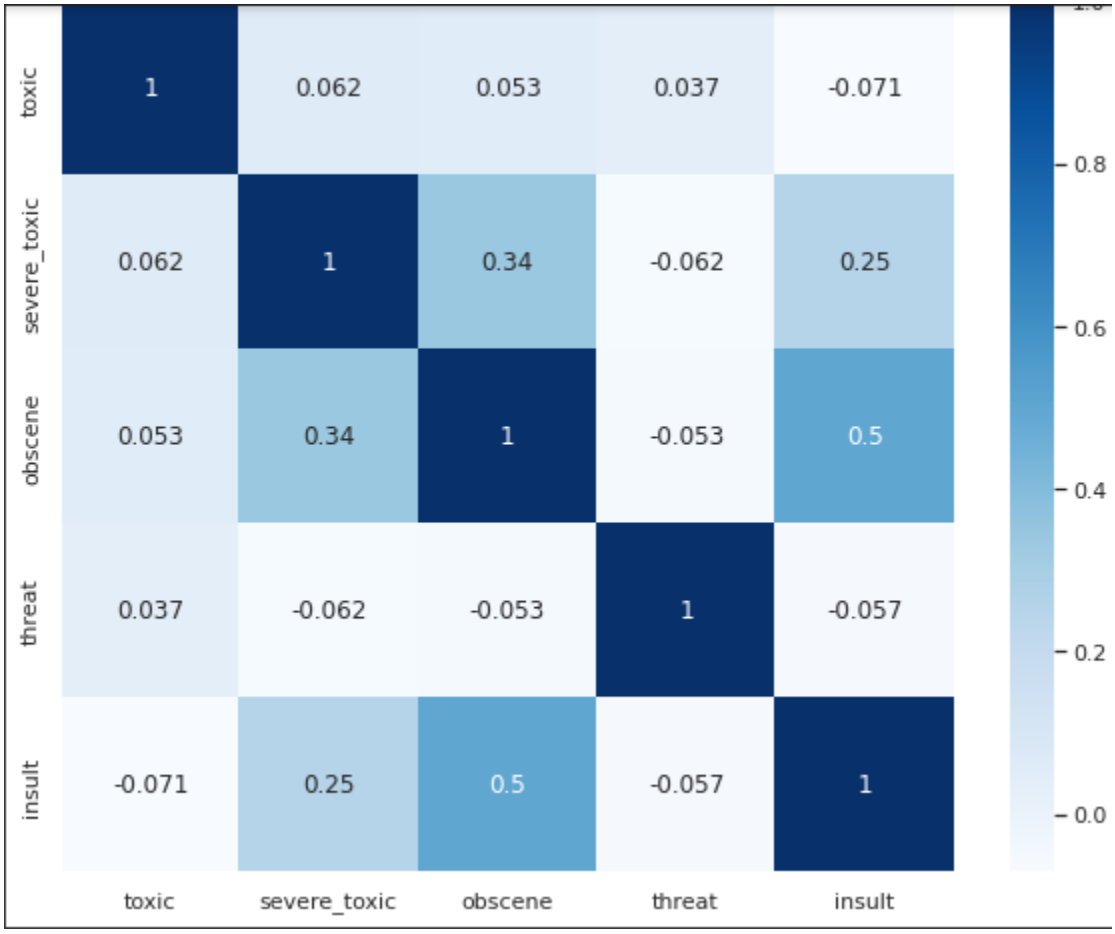
```
***** Running Evaluation *****
  Num examples = 84
  Batch size = 8
saved model
***** Running Evaluation *****
  Num examples = 84
  Batch size = 8
saved model
***** Running Evaluation *****
  Num examples = 84
  Batch size = 8
saved model
***** Running Evaluation *****
  Num examples = 84
  Batch size = 8
saved model
***** Running Evaluation *****
  Num examples = 84
```

بخش دوم:
مطابق خواسته سوال، داده های خود را با نسبت 80 و 20 برای آموزش و تست و به صورت متناظربرای کلاس های گوناگون تقسیم کردیم. پس از load کردن دیتا،ابتدا از طریق کد تحلیلی درباره نحوه پراکندگی داده های هر label داشتیم که نتیجه آن در شکل زیر آمده است:



همانطور که مشخص است، داده های toxic بیشترین تعداد را به خود اختصاص داده اند.

تحلیل دیگر ما درباره اشتراک و ارتباط بین label های مختلف بوده است. نتیجه این تحلیل در شکل زیر قابل مشاهده است.



همانطور که مشخص است، obscene و insult اشتراک بالایی با یکدیگر دارند. یعنی تعداد قابل توجهی از داده های insult، obscene را هم شامل می شوند و برعکس.

سپس، یک تابع مخصوص توکنایز کردن طراحی شده است. همچنین از مدل TFID بهره گرفته ایم تا تاثیر منفی داده های پرتکرار کمتر شود.
نهایتا از سه classifier زیر برای classification استفاده کرده و نتایج را ثبت کرده ایم:

```
MultinomialNB
LogisticRegression
LinearSVC
```

مقایسه نتایج به شرح زیر است:

	Model	Label	Recall	F1
0	MultinomialNB	toxic	0.125000	0.196768
1	MultinomialNB	severe_toxic	0.000000	0.000000
2	MultinomialNB	obscene	0.100000	0.160000
3	MultinomialNB	threat	0.000000	0.000000
4	MultinomialNB	insult	0.050000	0.080000
5	MultinomialNB	identity_hate	0.000000	0.000000
6	LogisticRegression	toxic	0.075000	0.124444
7	LogisticRegression	severe_toxic	0.000000	0.000000
8	LogisticRegression	obscene	0.100000	0.160000
9	LogisticRegression	threat	0.000000	0.000000
10	LogisticRegression	insult	0.050000	0.080000
11	LogisticRegression	identity_hate	0.000000	0.000000
12	LinearSVC	toxic	0.209722	0.275608
13	LinearSVC	severe_toxic	0.000000	0.000000
14	LinearSVC	obscene	0.300000	0.396190
15	LinearSVC	threat	0.000000	0.000000
16	LinearSVC	insult	0.250000	0.313333
17	LinearSVC	identity_hate	0.000000	0.000000

همانطور که مشخص است، LinearSVC Classifier برتری نسبی دارد.

بخش سوم:
ابتدا در کلاس commentDataset در تابع data read دیتای خود را برای هر برچسب به بخش train و test با نسبت 80 به 20 تقسیم کردیم
برای بهبود مدل خود در این بخش از مدل پیچیده تر Bert (transformer) استفاده کردیم و داده را در تابع tokenize_dataset توکن کردیم و input_ids و attention_mask را در تابع getitem استفاده می کنیم و سپس در کلاس CommentModel علاوه بر Bert از سیگموید، یک لایه Linear و نهایتا یک نورون استفاده شده است که تصمیم گیرنده ی این مسئله ی Binary Classification است.
و سپس مدل را train کرده و در تابع evaluate محاسبه ی accuracy و perdiction را انجام دادیم.

```
done 78 from 100
done 79 from 100
done 80 from 100
done 81 from 100
done 82 from 100
done 83 from 100
done 84 from 100
done 85 from 100
done 86 from 100
done 87 from 100
done 88 from 100
done 89 from 100
done 90 from 100
done 91 from 100
done 92 from 100
done 93 from 100
done 94 from 100
done 95 from 100
done 96 from 100
done 97 from 100
done 98 from 100
done 99 from 100
done 100 from 100
88.5
```