بهنام ساعدی ۴۰۱۲۰۲۹۴۲ حامد سعادتی ۴۰۱۲۱۲۸۸۴

فاز ۱ پروژه ی MlOps

در این فاز دیتاست های مختلف بررسی شده و فعلا برای این فاز یک دیتاست انتخاب شده است که پیش پردازش های کلی روی آن اعمال شود و در فاز های بعدی امکان اضافه کردن دیتاست های دیگر نیز وجود دارد.

دیتاست مورد استفاده :

Dataset

The following dataset will be used as training data

- Hammer, H. L., Riegler, M. A., Øvrelid, L. & Veldal, E. (2019). "THREAT: A Large Annotated Corpus for Detection of Violent Threats". 7th IEEE International Workshop on Content-Based Multimedia Indexing.

This dataset consists of a total of around 30,000 sentences from around 10,000 YouTube comments. Each sentence is manually annotated as either being a violent threat or not

داده های دیتاست در فایل txt. ضمیمه شده است.

ابتدا پیش پردازش هایی مثل پاکسازی عبارات متنی با استفاده از عبارات منظم انجام شده است (کارهایی مثل حذف نشان های نگارشی و حذف duplicates و ...)

سپس histogram طول متن های ورودی plot شده است و با توجه به آن maximum sequence length دیتاست انتخاب شده است و داده هایی که طولی بیشتر از این maximum length دارند ، حذف شده اند.

سپس با استافده از nltk عملیات stemming و lemmatization انجام شده است.

در بخش sampling ابتدا stratified sampling بررسی شده است که توزیع کلاس های مختلف را هنگام split کردن در داده های test , train حفظ می کند.

Sampling استفاده شده در این قسمت balanced sampling می باشد و برای تقسیم آن به بخش های train و test ، از stratified sampling استفاده شده است.

معیار ارزیابی ما در این جا معیار f1_score می باشد که با معیار بسیار خوب و دقیقی می باشد .

در بخش بعدی برای یافتن representation متون دیتاست از رویکر tf_idf و pretrained bert model استفاده شده است و برای این قسمت با استفاده از SVD ، کاهش بعد نیز انجام شده است و ابعاد ماتریس tf_idf از حدود ۲۰۰۰ به ۲۰۰ رسیده است که ۱/۱۰ برابر شده است.

و در ادامه یک مدل multinomial naïve bayes روی داده ها fit شده است که برای هر دو کلاس دقتی حدود ۸۸-۸۹ درصد را داده است.

توضیحات بیشتر همراه با نتایج اجرا در فایل ipynb. ضمیمه شده است.