Project: Hate Speech detection using Transformers (Deep Learning)

Group Name: ltvn_sergey
Name: Litvinov Sergey
Email: ltvn.sergey.work@gmail.com
Country: Russia
Company: NVI Solutions
Specialization: NLP

Problem description

Hate speech detection is a problem of sentiment classification. The task is to classify certain piece of text as the one with hate speech or not, by training deep learning model on labeled data.

The term hate speech is understood as any type of verbal, written or behavioral communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, color, ancestry, sex or another identity factor.

Data understanding

There are two csv format datasets in available:

- **train.csv** - with id, tweet with text and label - 1 for hate-speech and 0 - otherwise
- **test.csv** - with id, tweet with text

Data problems and possible solutions

1. **Class imbalance**
   Data has significant class imbalance:
   - 93% of tweets are positive
   - 7% of tweets are negative

   Imbalance in data could lead to bias in our model
   To counter imbalance during training we can use following techniques:
   1. Downsample the majority class
   2. Upsample the minority class
   3. Use weights for the classes at the algorithmic level
   4. Use stratified cross-validation

2. **Text cleaning**
   Text data is noise and has a lot of extra symbols, uninformative signs, misspelled words.
   For cleaning text data we will perform following steps:
   1. Remove symbols, digits, extra spaces using regular expressions
   2. Remove URLs using regular expressions
   3. Normalize casing
   4. Tokenize sentences into individual terms, using tokenizers
   5. Remove stop words