NLP Project phase 1

Saba Rayehi - Sohrab Namazi Nia
Iran University of Science and Technology

Toxic comment definition

The goal of this project is to classify comments into two groups, toxic comments, and safe comments. A comment must be classified as toxic if it is rude, disrespectful, or unreasonable.

Toxic comment Dataset

We searched for many different datasets on the internet and finally chose one of them that was a better generalization of real-world comments.

The raw dataset can be found here:

Link

Dataset Structure

The raw dataset is written into CSV files. Each element of the training dataset, consists of a clear label, stating whether the data is toxic or not. Also, there are 5 parameters for each element that are mentioned here:

Severe_toxic

Obscene

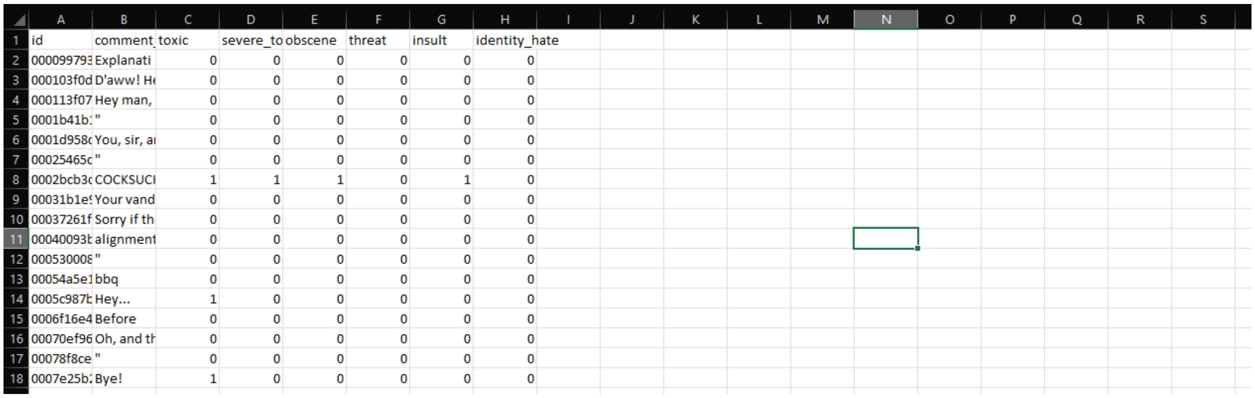
Threat

Insult

Identity_hate

The values for the mentioned parameters are binary, meaning that for example, a comment is whether a threat or not. Also, each comment has a unique ID.

You can see a demo of what the CSV file for training data looks like in the following figure:



id	comment_	toxic	severe_to	obscene	threat	insult	identity_hate
00009979	Explanati	0	0	0	0	0	0
000103f0d	D'aww! H	0	0	0	0	0	0
000113f07	Hey man,	0	0	0	0	0	0
0001b41b	"	0	0	0	0	0	0
0001d958	You, sir, a	0	0	0	0	0	0
00025465	"	0	0	0	0	0	0
0002bcb3	COCKSUC	1	1	1	0	1	0
00031b1e	Your vand	0	0	0	0	0	0
00037261f	Sorry if th	0	0	0	0	0	0
00040093	alignment	0	0	0	0	0	0
00053000	"	0	0	0	0	0	0
00054a5e	bbq	0	0	0	0	0	0
0005c987	Hey...	1	0	0	0	0	0
0006f16e4	Before	0	0	0	0	0	0
00070ef96	Oh, and th	0	0	0	0	0	0
00078f8ce	"	0	0	0	0	0	0
0007e25b	Bye!	1	0	0	0	0	0

Dataset statistics

We have acquired different asked parameters about the dataset in our code. For example sentence count, word count, etc.

We use nltk and its functions :

For separation sentences —-> we use function sent_tokenize

For separation words—--> we use function word_tokenize

For clean _data:

To clean the data, we first removed the spaces and stopwords then the nonsense words and characters