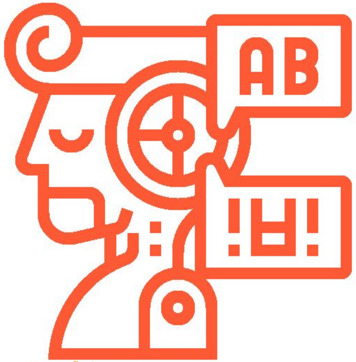
The Detoxifiers



https://www.flaticon.com/free-icons/natural-language-processing

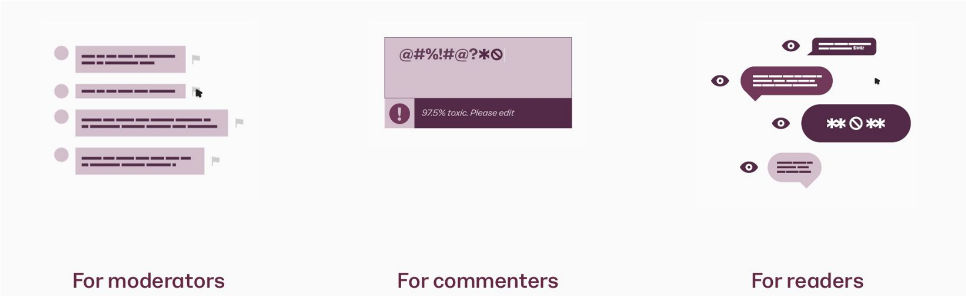
Empower Fruitful Discussions

Ksenia Gerasimovich, Katharina Neumüller, Daniel Guhr & Kristin Stöcker

Toxicity online

- Immense number of comments every day
- Toxic trolls who conquer the discussion boards hinder online discussions

- The threat of abuse and harassment online → many people stop expressing themselves
- Content moderators are not being able to moderate all of it anymore
- Disabling of discussions due to high cost
- Many communities tend to limit or completely shut down user comments
- Platforms struggle to effectively facilitate conversations

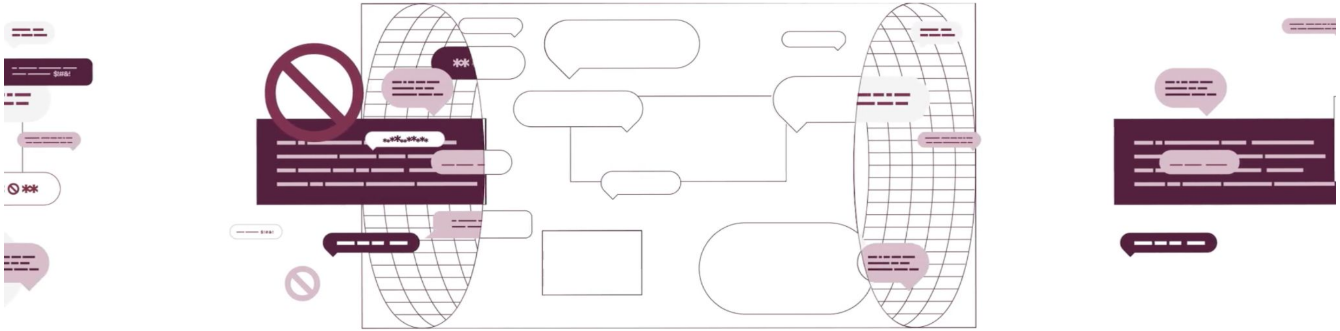


Source: https://perspectiveapi.com

Toxicity online

Machine learning algorithms are used to reduce toxicity online:

- The Conversation AI team are working on tools to help improve online conversation
- Area of focus: study of negative online behaviors, like toxic comments
- Perspective API: publicly available models to moderate a content and restrict toxicity
- These models still makes erros



Source: https://perspectiveapi.com

Challenge description

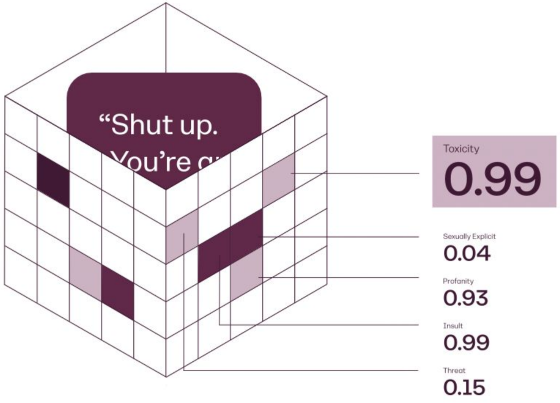
Task

- to build a multi-headed model that's capable of detecting different types of of toxicity
- that operates better than Perspective's current models

Dataset:
- comments from Wikipedia's talk page edits

Train dataset:
- a number of Wikipedia comments, labeled by human raters for toxic behavior

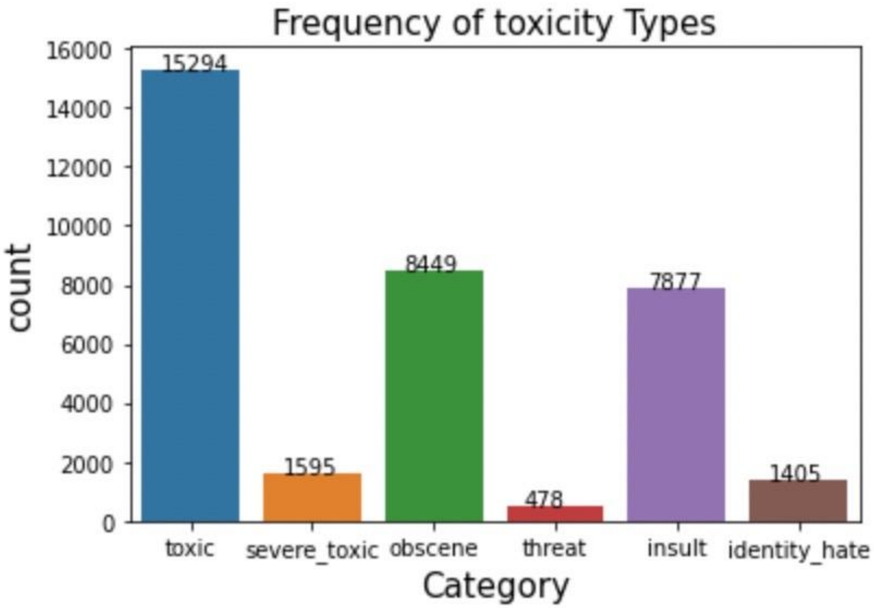
Evaluation metric:
- the mean column-wise ROC AUC



Source: https://perspectiveapi.com/how-it-works/

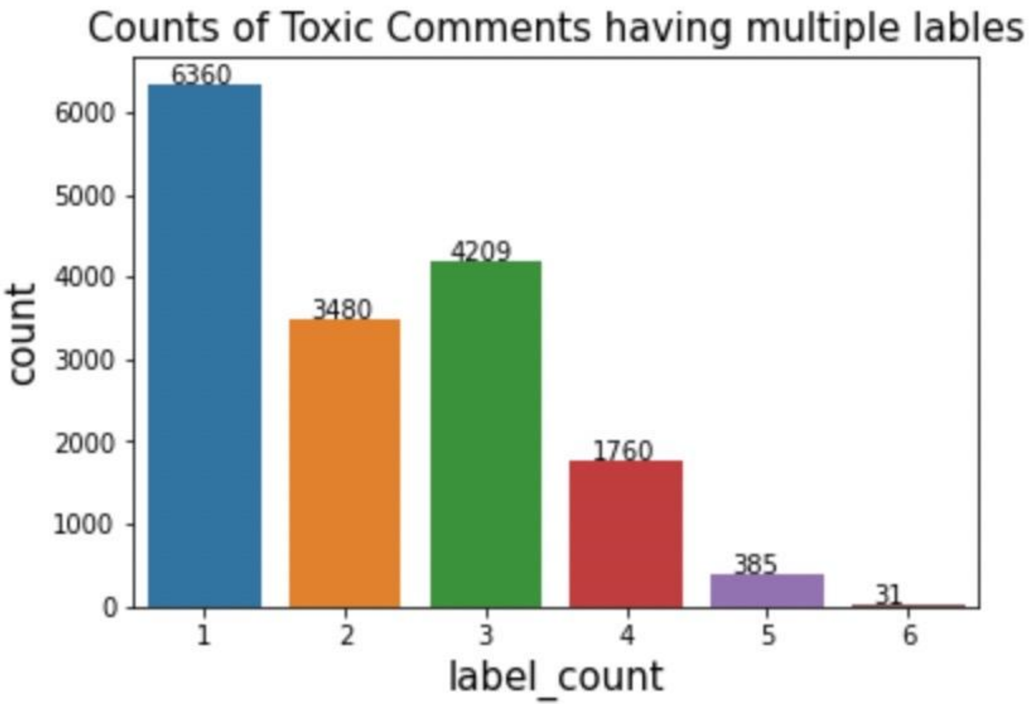
Dataset

- The types of toxicity are:
  - **toxic**
  - **severe_toxic**
  - **obscene**
  - **threat**
  - **insult**
  - **identity_hate**
- **159.571** samples
- **~10%** are toxic comments



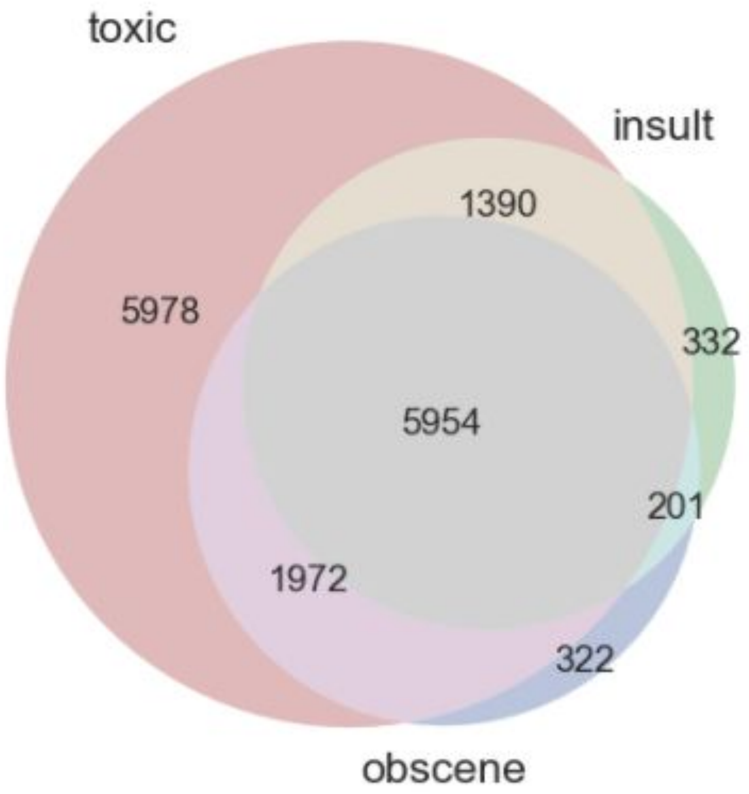
Toxic Comments with multiple labels

- **Multilabel Classification Problem**
- Many toxic comments have multiple labels
- Most of the comments have 2 - 3 labels
- Only a few comments have more than 4 labels



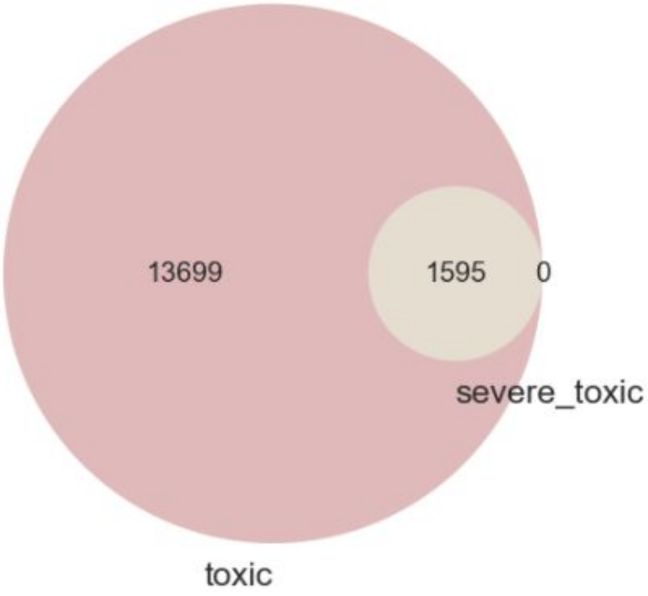
Venn Diagram for ‘toxic’, ‘insult’ and ‘obscene’

- Many of the toxic comments are also labeled as **insult** or **obscene**
- only a small part of **obscene** and **insult** that are not also labeled as **toxic**
- **5954** comments have all three labels



Venn Diagram for ‘toxic’ and ‘severe toxic’

- The category **severe_toxic** is contained in **toxic**
- There is a semantic link between the two category names
- **severe_toxic** represents just 11.64% of toxic comments



Toxic

Insult

Obscene

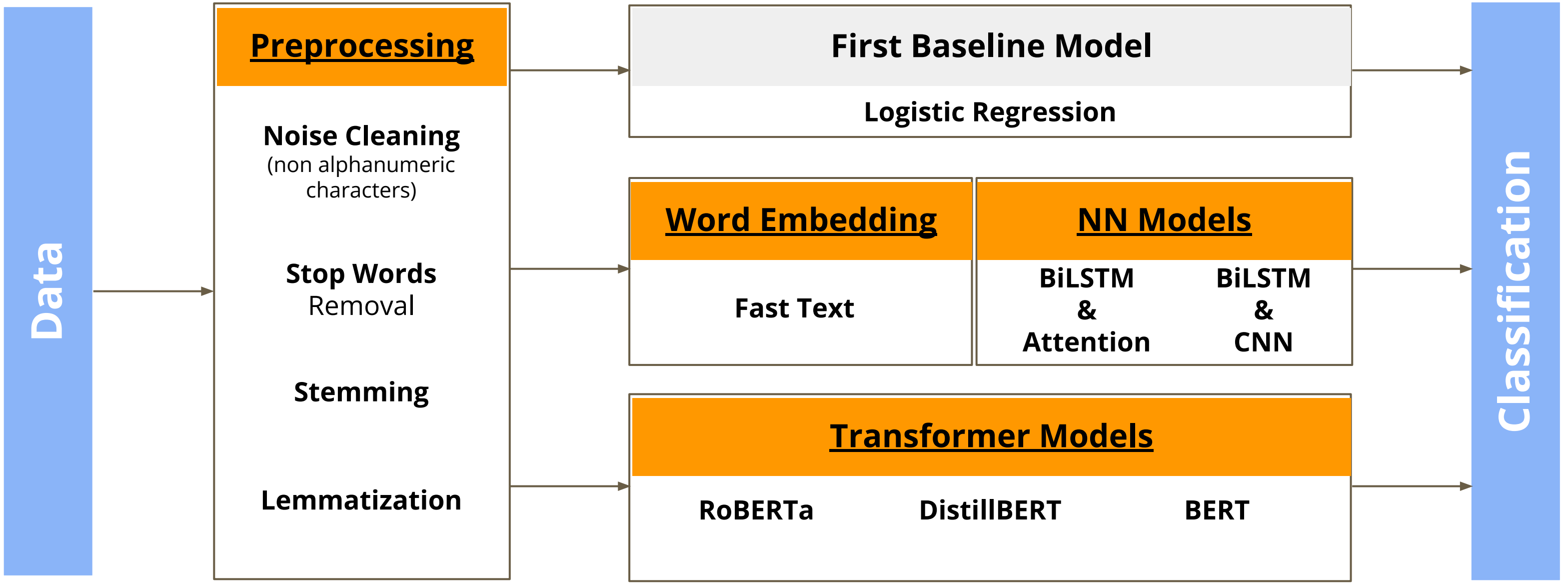
Threat

Severe Toxic

Identity Hate

Model Architecture

A modular approach



Preliminary Results

- At this point of the project **BERT** is the undisputed front runner
- pretrained model from huggingface.co fitted on the data set
- almost no preprocessing on given data set
- ROC-AUC score:

$\boxed{0.98433}$

Preliminary Results

→ our score ranks us on position 1560 of 4539 on the official leaderboard on kaggle.com

→ distance to first place 0.00486 (a score of 0.98901)

1558	▲ 15	Franck My	0.98434	2	4Y
1559	▲ 15	Belette Chaton	0.98434	7	4Y
1560	▲ 24	The Detoxifiers	0.98433	3	3D
1561	▲ 24	Артем Лян	0.98433	19	4Y
1562	▼ 29	MSJose	0.98432	30	4Y

Preliminary Results

Class Labels:

Toxic 0
Severe Toxic 1
Obscene 2
Threat 3
Insult 4
Identity Hate 5

Classification Report :

	precision	recall	f1-score	support
0	0.83	0.84	0.84	4591
1	0.56	0.39	0.46	485
2	0.82	0.86	0.84	2527
3	0.56	0.51	0.53	131
4	0.75	0.79	0.77	2362
5	0.63	0.51	0.56	430
micro avg	0.79	0.80	0.79	10526
macro avg	0.69	0.65	0.67	10526
weighted avg	0.79	0.80	0.79	10526
samples avg	0.07	0.08	0.07	10526

F1-Score:

- a weighted metric to measure the model's performance
- gives an idea of performance in different classes

→ model performs worse on less frequent classes

Outlook

- How do we get better?

Outlook

- How do we get better?
    - **Different models**
        - Comparing results of BERT to RoBERTa and XLNet
        - Using datasets with different preprocessing

Outlook

- How do we get better?
  - **Different models**
    - Comparing results of BERT to RoBERTa and XLNet
    - Using datasets with different preprocessing
  - **Data augmentation**
    - Over-/Undersampling
    - Oversampling of smaller classes by using translations

Outlook

- How do we get better?
  - **Different models**
    - Comparing results of BERT to RoBERTa and XLNet
    - Using datasets with different preprocessing
  - **Data augmentation**
    - Over-/Undersampling
    - Oversampling of smaller classes by using translations
  - **Error analysis**
    - Where does our model make mistakes?
    - Is there anything we can do to counter these mistakes?

Outlook

- How do we get better?
    - **Different models**
        - Comparing results of BERT to RoBERTa and XLNet
        - Using datasets with different preprocessing
    - **Data augmentation**
        - Over-/Undersampling
        - Oversampling of smaller classes by using translations
    - **Error analysis**
        - Where does our model make mistakes?
        - Is there anything we can do to counter these mistakes?
    - **Fine tuning**
        - Gridsearch or Keras Tuner to fine tune hyperparameters

Thank you for your Attention!

Any Questions?

Ksenia Gerasimovich, Katharina Neumüller, Daniel Guhr & Kristin Stöcker

Sources and Libraries

Wordcloud Library

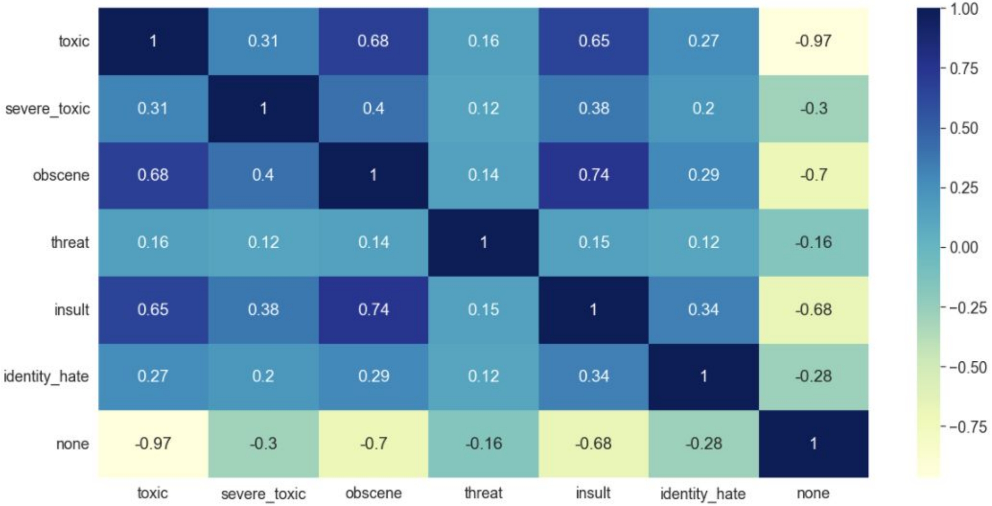
https://github.com/amueller/word_cloud

Logo

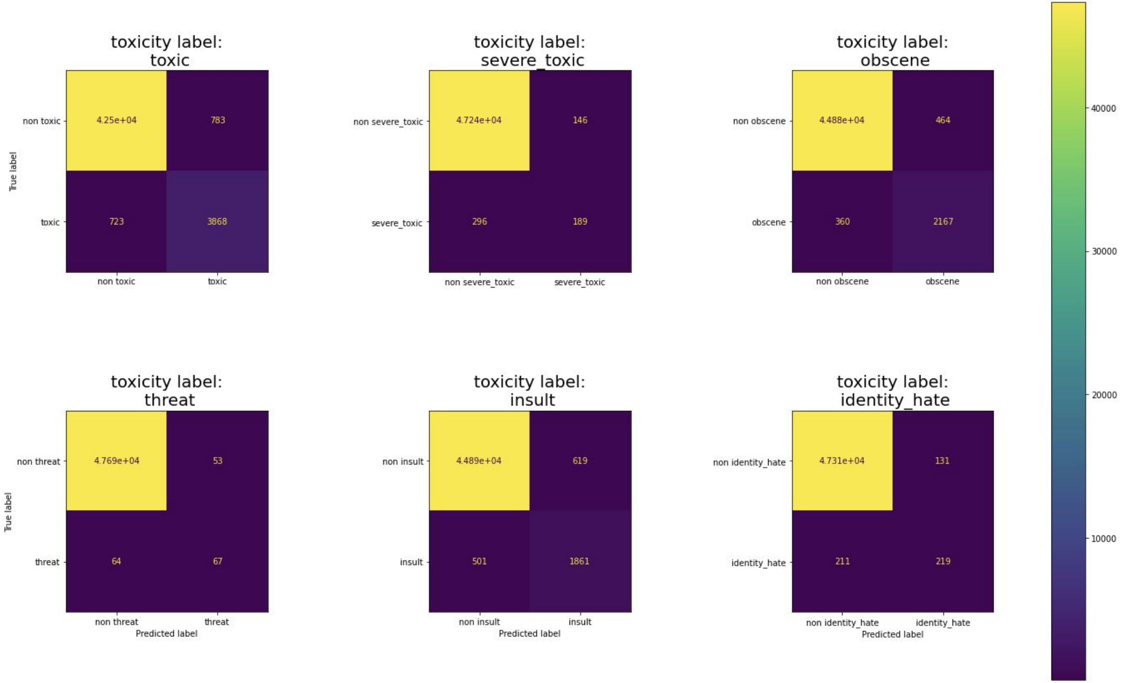
Natural language processing icons created by Eucalyp - Flaticon

Correlation between the class labels

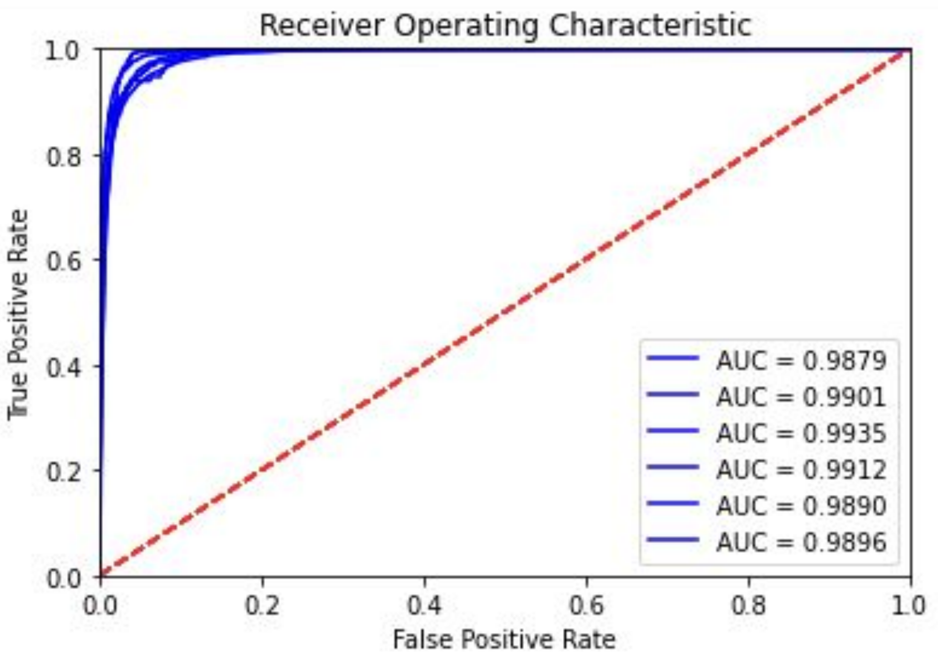
- There is a strong relation between the labels
  - **toxic** and **insult**
  - **toxic** and **obscene**
  - **obscene** and **insult**
- In the presence of one of the labels in the pairs for a comment, it is likely that the comment will also have the other label



BERT: Confusion Matrix



BERT: ROC AUC



- ROC: **R**eceiver **O**perator **C**haracteristic
- AUC: **A**rea **U**nder the **C**urve
- way of visualizing the performance of a model, plotting true positives against false positives
- the bigger the area under the curve the higher the score and the more true positives and less true negatives

BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Is backed by Transformer and it’s core principle - attention, which understands the contextual relationship between different words
- Was pre-trained on unsupervised Wikipedia and Bookcorpus datasets using language modeling
- Learns information from a sequence of words not only from left to right, but also from right to left

Self-Attention

- We can think **“self-attention”** means the sentence will look at itself to determine how to represent each token
- When the model processes the word **it,** the self-attention looks at other words for better encodings
- One word can have different meanings in different sentences (**context**), and self-attention can encode (**understand**) each word based on context words for the current word.