Jigsaw Multilingual Toxic Comment Classification

Aashish Acharya, Cosima Birkmaier, Philip Bramwell

opencampus.sh

24W | Intermediate Machine Learning

Trigger Warning & Disclaimer

The project deals with toxic language. The presentation may therefore (for purely academic purposes) contain toxic - particularly offensive, obscene, threatening or hateful content. It is not our intention to offend anyone in the audience or anyone's mother.

Objective

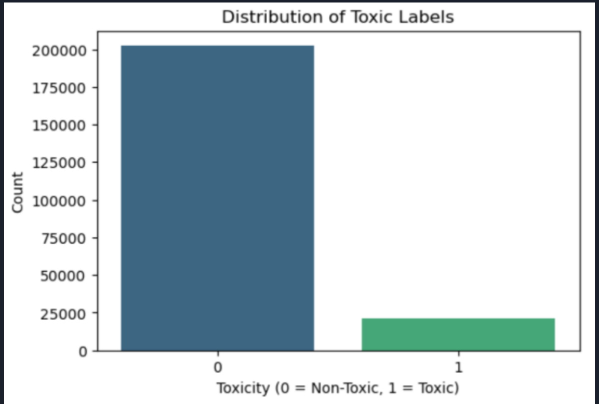
- Predict the probability of a comment text being toxic
- “Toxic” includes:
    - Hate speech
    - Harassment
    - Insults
    - Abusive language

Dataset

- Comments from Wikipedia talk pages & Civil Comments platform
- Training set:
  - 223,549 samples
  - 8 columns: id, comment, toxic, severe_toxic, obscene, threat, insult, identity_hate
  - Label: column 'toxic'
  - All comments in English
- Validation set:
  - 8,000 samples
  - Comments in 3 languages (Turkish, Spanish, Italian)
- Test set:
  - 63,812 samples
  - Comments in 6 languages (Turkish, Spanish, Italian, Russian, Portuguese, French)

Challenge: Imbalance of Class Distribution

- Class distribution:
  - 202,165 (90.4%) non-toxic
  - 21,384 (9.6%) toxic
- Techniques to address this:
  - Undersampling
  - Random Oversampling
  - Data Augmentation:
    - Translation & Back-Translation
    - Contextual augmentation
  - ROC-AUC metric (Receiver Operating Characteristic - Area Under the Curve)

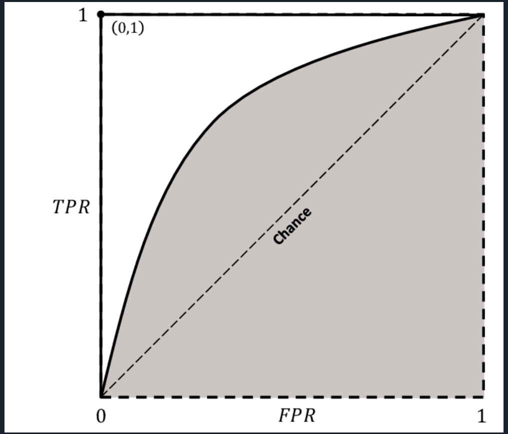


Challenge: Different Sets of Languages

- Training set:
  - English
- Validation set
  - Turkish, Spanish, Italian
- Test set
  - Turkish, Spanish, Italian, Russian, Portuguese, French
- Methods to address this:
  - Use of translations
  - Use XML-RoBERTa model
    - Pre-trained on >100 languages
    - Recognizes languages automatically
    - Preferred by most successful participators of the Kaggle competition

ROC-AUC

- ROC Curve:
  - Plots True Positive Rate (TPR) vs. False Positive Rate (FPR) across Classification thresholds.
- AUC:
  - 1: Perfect separation.
  - 0.5: Random guessing (diagonal line).
- Why use it?:
  - Threshold-independent performance metric.
  - Robust for imbalance datasets.



https://www.researchgate.net/figure/ROC-curve-with-AUC-shaded-in-gray_fig7_367283054

Contextual Augmentation

- Definition and core concept
  - Generates synthetic training samples by leveraging contextual understanding
  - Focuses on preserving semantic meaning while altering text
- Implementation
  - Bidirectional Encoder (BERT) Model
  - Probability of 60%
  - Maximal 10 word substitutions in a sequence
  - Handling imbalanced dataset
  - Generating only toxic comments

```
Original counts - Toxic: 4883, Non-toxic: 45117
```

```
Train size: 89064, Val size: 8000, Test size: 63812
```

Translation and Back-Translation

- Translate toxic training comments from English to 6 different languages
- Translate each translation back to English
- Implementation:
  - deep_translator GoogleTranslator
  - Accesses GoogleTranslator online
  - Very time consuming
  - Susceptible to failure
  - 12,636 original toxic training samples were used
  - 151.632 new, augmented samples

Your are a filthy fuck go fuck your self

Sen pis bir orospusun git kendini becer	You're a dirty whore go fuck yourself
Ты грязный ублюдок иди нахуй	You are a dirty bastard go fuck yourself
Sei un fottuto stronzo vai a farti fottere	You're a fucking asshole Go fuck yourself
Tu es un sale connard va te faire foutre	You're a dirty asshole fuck you
Você é um filho da puta imundo vai se foder	You're a filthy son of a bitch fuck you
Eres un asqueroso cabrón vete a la mierda	You are a disgusting bastard go to hell

Baseline Model: Architecture

RNN Baseline Model

- Embedding Layer: Maps words/tokens to 100 dimensional vectors
- RNN Layer: 128 hidden units
- Classification Head: Sigmoid output for binary prediction

Training Details:

- Loss: BCEWithLogitsLoss
- Optimizer: Adam (lr=0.001)
- Epochs: 5

Evaluation:

- Accuracy on test set/split: 90.40% (trained on reduced dataset of 10,000 samples)
- Test AUC: 0.5275

DistilBERT

- Trained on small subset of 5,000 samples
- Tokenized data with DistilBERTs own tokenizer
- 1 epoch, batch size = 8, learning rate = 2e-5
- ROC-AUC: 0.7287

XLM-RoBERTa with Undersampling

- Original dataset size: 223,549
- Balanced dataset size: 42,768
- Toxic comments: 21,384
- Non-toxic comments: 21,384
- Filtering comments exceeding max token limit 512
- 1072 comments removed from training dataset
- 2 epochs, batch size = 8, learning rate = 2e-5

Epoch	Training Loss	Validation Loss	Accuracy	Roc Auc
1	0.190200	0.330726	0.873000	0.911624
2	0.180200	0.461748	0.868625	0.906466

XLM-RoBERTa with Contextual Augmentation

- Trained for 3 epochs
- Potentially overfitting to the training data
- ROC-AUC as metrics
    a. Measures the model's ability to discriminate between classes across all thresholds
    b. Reflects overall ranking performance

Epoch	Training Loss	Validation Loss	Roc Auc
1	0.199000	1.787323	0.903857
2	0.180100	1.429178	0.915215
3	0.149500	1.606748	0.914398

- ROC-AUC of test set: 0.8934

XLM-RoBERTa with Translations & Back-Translations

- All augmented comments
    - Large dataset with 375,181 samples
    - 2 epochs, batch size = 128, learning rate = 2e-5
    - ROC-AUC: 0.7580
    - Problems with overfitting
- Only translations & undersampling
    - 194,400 samples
    - 2 epochs, batch size = 32, learning rate = 3e-5
    - ROC-AUC: 0.7687
- Only back-translations & undersampling
    - 194,400 samples
    - 2 epochs, batch size = 32, learning rate = 3e-5
    - ROC-AUC: 0.8269

Comparison

Model	Augmentation / Address imbalance	ROC-AUC
Baseline (RNN)	-	0.5275
DistilBERT	-	0.7287
XLM-RoBERTa	Undersampling	0.9116
XLM-RoBERTa	Translation & Back-Translation	0.7580
XLM-RoBERTa	Translation & undersampling	0.7687
XLM-RoBERTa	Back-Translation & undersampling	0.8269
XLM-RoBERTa	Contextual Augmentation	0.8934

Summary

- Results are not satisfying so far
- Some data augmentation methods (like translation) seem to decrease performance
- Undersampling was so far the best performing method

Sources

Kobayashi, S. (2018): Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201

Shleifer, S. (2019): Low Resource Text Classification with ULMFit and Backtranslation. https://arxiv.org/abs/1903.09244

Wu, X. & Lv, S. & Zang, L. & Han, J. & Hu, S. (2019). Conditional BERT Contextual Augmentation. 10.1007/978-3-030-22747-0_7