

# Better toxic language classification despite data scarcity

*What kind of **data augmentation** is effective for **toxic language classification**?*

## Problem:

- Dataset labeling is **expensive** → augment with novel **synthetic samples**
- Effect on **toxic language classification** not studied before

## Our contributions:

- Applying data augmentation on Kaggle's toxic language dataset (**threat** label)
- Comparison of **8 augmentation techniques** across **4 classifiers**

## Challenges in toxic language data:

- **Small seed datasets**
- **High class imbalance**: non-toxic data more common and easier to obtain

## Most effective techniques (up to **21% improvement** in F1-score):

- Replacing subword tokens with **BytePair Embedding neighbours**
- Adding **random non-toxic sentences** to original toxic training documents
- Generating new toxic documents with the **GPT-2** language model