1 Datamining

Primary sources of the information were:

1. ParaNMT filtered corpus [**?**]
2. Glove embeddings [**?**]
3. WordNet [**?**]
4. Jigsaw bad words dataset [**?**]

1.1 ParaNMT

While the dataset was considered as a dataset of the toxic sentences and their non-toxic paraphrases with toxicity ranking from people, it seems to be slightly wrong.

It seems that paraphasing is the reverse-translation of the sentence, which was done artificially as well.

Moreover, original work [**?**] for toxicity ranking proposed one more model.

As a result, the main dataset was artificially generated, artificially ranked and in general is not a very good source of data.

1.1.1 Filtering the filtered

To achieve decent results we can carefully pick the samples from the provided dataset. For instance, experimentally we deduced that filtering the samples which lead from high level of toxicity to the very low in result, filters many insane examples.

Moreover, such filter is beneficial in terms of actually successfully detoxifying the text, while preserving the high similarity of the reference and translation.

Final parameters for filtering:

$$\text{tox}_{\text{ref}} > 0.99 \wedge \text{tox}_{\text{trn}} < 0.01 \quad (1)$$

1.1.2 Splitting

After filtering the data, we can deterministically sample the data with specified random seed, to achieve reproducible results. For instance, we want 3000 samples for test data. 10000 of samples will be split in $9:1$ correspondence for train/validation datasets. While larger samples are computationally feasible, they tend not to improve final results greatly, as a result, we will stick with these values.

2 Metrics

2.1 Toxicity

To objectively compare the performance of the models in text detoxification we need some toxicity metric. As such, we are going to use ***Detoxify*** package [**?**].

Each model predictions will be ranked in the interval $[0, 1]$ for further comparison.

2.1.1 Semantic similarity

Since we have a task of translation (style change) with preservation of the meaning, we might want to use some related metrics. Generally, we have decided to stick with metrics introduced in initial draft.

Final results will be compared using the following metrics:

- BLEU
- METEOR
- Cosine Similarity (GLoVe embedded)