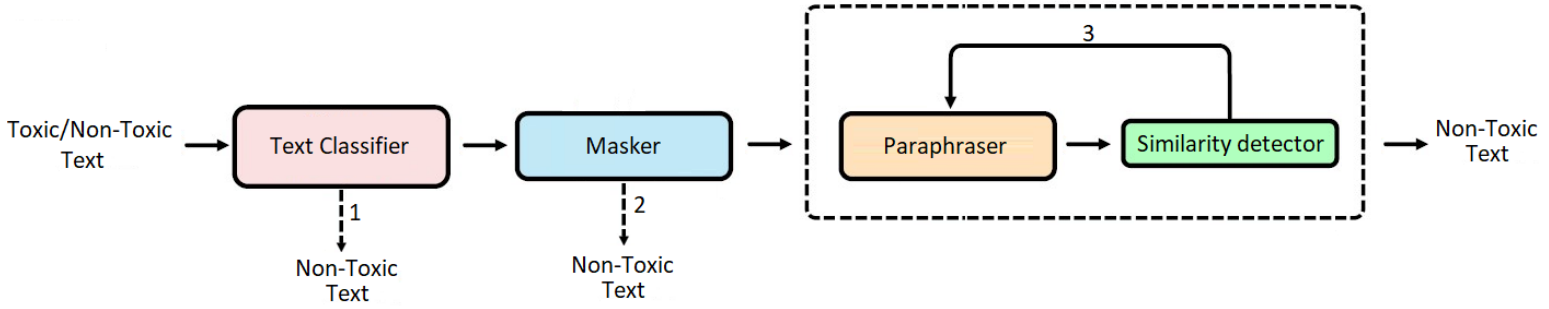
Report. Final model.

Final model has the following architecture:



Main parts:

1. Text Classifier:

   Based on the pre-trained zero-shot classification facebook/bart-large-mnli model. Binary classifier, which assigns a label to the text: toxic/non-toxic. In case of non-toxic text, it returns it back (1 at figure). Toxic text goes further to Masker.

2. Masker:

   Based on a custom seq2seq model with LSTM cells. Masker tokenize text and assign binary label - toxic/non-toxic - to each non-punctuation token. For an empty list of toxic tokens, text is considered as non-toxic and returns back (2 at figure). For toxic text Masker hides each toxic token separately, which means that it produces a list of sentences, where each sentence contains exactly one masked word. This list goes to Full-mask detoxifier.

3. Full-mask detoxifier:

   Performs replacement of masked tokens by non-toxic words.

   3.1. Paraphraser:

   Based on the pre-trained distilbert-base-uncased model, which generates candidate words for replacement. However, sometimes proposed words are toxic (even more toxic than the initial word), so it uses the first part of Masker to find several non-toxic candidate words.

   3.2. Similarity detector:

   Based on distilbert-base-nli-mean-tokens model, which checks the semantic similarity of sentence with proposed candidate and initial sentence, to obtain the most appropriate word (3 at the figure).

As a result, we obtain non-toxic text.