Supplementary material

Anonymous NAACL-HLT 2021 submission

1 Datasets used

We make use of the following datasets in our experiments to demonstrate the effectiveness of the attack algorithms:

1. We make use of the dataset (Modified ADE Corpus) [1], ie. the ADE (adverse drug effect) corpus, was created by (Gurulingappa et al., 2012b) by sampling from MEDLINE case reports. Each case report provides important information about symptoms, signs, diagnosis, treatment and follow-up of individual patients. The ADE corpus contains 2, 972 documents with 20, 967 sentences. Out of which, 4, 272 sentences are annotated with names and relationships between drugs, adverse effects and dosages. It does not contain a fixed training and testing corpus. Thus we make use of 10-fold cross-validation in our experiments and report results in the original paper.

2. Another one is a Twitter dataset (Sarker et al., 2016) published for a shared task in Pacific Symposium on Biocomputing, Hawaii 2016. The tweets associated with the data were collected using generic and brand names of the drugs, and also their possible phonetic misspellings. The tweets were annotated for presence of ADRs. In the shared task, 70% (7, 575) of the original data set is shared for training and the rest of the data is used for evaluation. Due to Twitter's data terms and conditions, only the tweet ids are contained in the original file. At the time of this experiment, we could download only 4, 974 tweets (with 498 tweets with ADR descriptions) as many tweets are no longer accessible.

2 Models used

We make use of the following models and their variants in our experiments. The huggingface transformers have been used for this purpose.

1. **Bert-based-uncased**:12-layer, 768-hidden, 12-heads, 110M parameters. Trained on lower-cased English text.

2. **Roberta-base**:12-layer, 768-hidden, 12-heads, 125M parameters RoBERTa using the BERT-base architecture

3. **Scibert-scivocab-uncased**: This model was pre-trained on the scientic articles from the Semantic Scholar.

4. **BioBERT-Base v1.0 (+ PubMed 200K + PMC 270K)** This is based on BERT-base-Cased. This model was pre-trained on large-scale biomedical corpora (PubMed and PMC articles).

5. **ClinicalBERT**: This model was trained on both all the clinical notes and discharge summaries.

[1] https://sites.google.com/site/adecorpus/home/document