Abstract

Natural language processing (NLP) has been an attractive research goal for many years, since solving this task in its general form will allow creating a natural language interface, which will greatly simplify and expand the scope of human computer interaction. To build a robust model, which can successfully solve the assigned task, data analysis must be performed carefully and thoroughly. Jigsaw's training set is quite large — about 2 000 000 comments. Each comment in the training set has a toxicity label — a target value, which models must be taught to predict.

In addition to the statistical characteristics of the training set, it is also necessary to consider characteristics of textual data. Comments are mainly written in English (the training set contains comments in other languages, but their number is less than 0.1% of the total). Many comments contain emojis, an excess of punctuation marks, web links, numbers, unusually written words, grammatical errors.

Methodology

Recurrent neural networks (RNNs) have proven themselves well to solve various NLP problems. The idea behind RNNs is to make use of sequential information. In a sense, an RNN "remember" the previous calculations and uses this information in the current processing. Such types of RNN as LSTM or GRU are best cope with the tasks of text classification, so when developing models to solve the problem of detecting toxic comments, it was decided to use them.

BERT (Bidirectional Encoder Representations from Transformers) is a new method of pre-training contextual language representation model developed by Google AI Language team. BERT was trained using only a large plain text corpus (like Wikipedia), which is important because an enormous amount of plain text data is publicly available on the web in many languages. It is able to obtain state-of-the-art results on a wide range of NLP tasks.

Taking into advantage the aforementioned toxic comment detection problem, it becomes clear that evaluation metric must be able to balance overall performance with various aspects of unintended bias.

The ensemble includes Bi-GRU-LSTM and Bi-GRU with attention mechanism models presented in this article with GloVe and FastText words embeddings and also 2 BERT models. Thus, 6 models were included in the ensemble. In addition, a technique called "seed averaging" was used (it was applied to Bi-GRU-LSTM and Bi-GRU with attention mechanism models), which consisted in launching one model with initializing a pseudorandom number generator with different values and averaging its predictions. To better illustrate the work of the presented

models, we test them on several sentences that contain references to some identities that most often suffer from toxic comments. As noted above, these sentences are complex for classical toxicity detection algorithms.

Conclusion

Deep learning technologies can minimize human participation in the development of algorithms, since creation of features specific to a particular task is automated. In this paper it was shown how to use deep NNs to solve the problem of detecting toxic comments. The results and accuracy of predictions obtained by the ensemble and each developed model separately, overwhelmed the results of the models from previously published works on this topic, which indicates the success of the work done. Further improvements in the accuracy of model predictions can be achieved using different augmentation techniques that increase the size of the training data set. Before each epoch during the training of a model, a subset of the comments from the training subsample has to be enriched with augmented data.