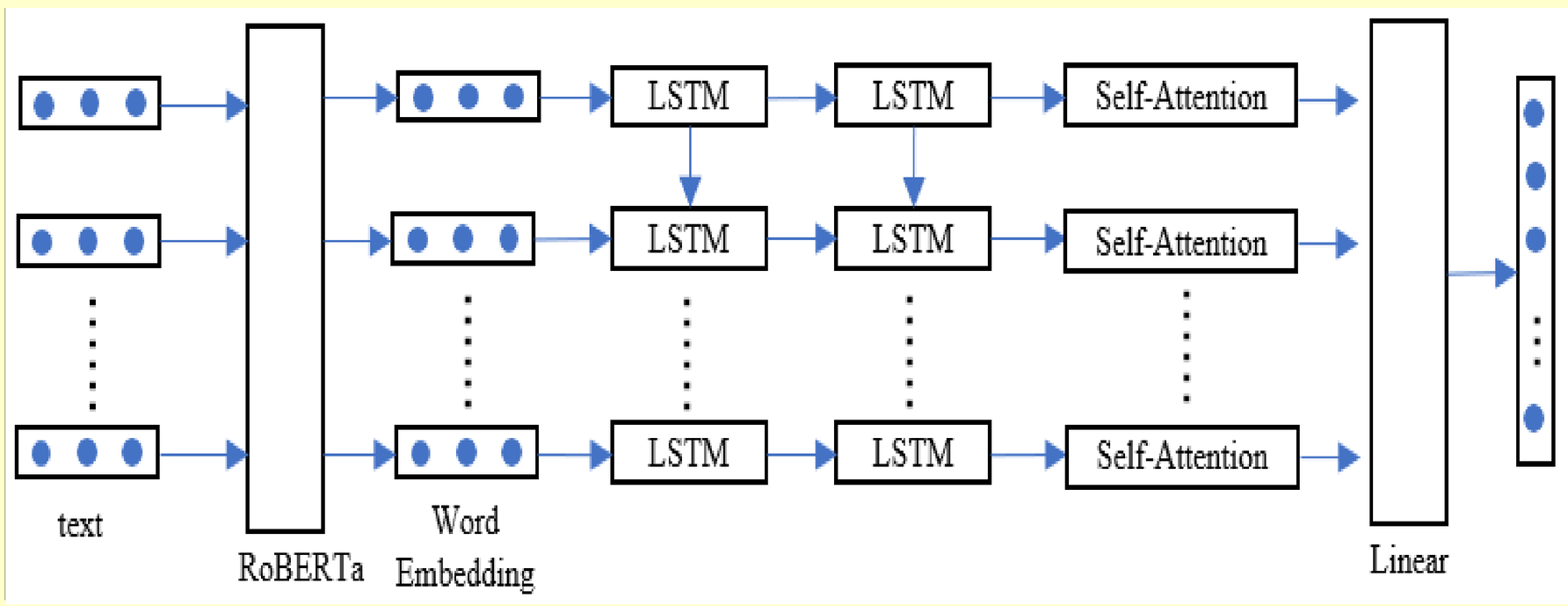
ICIEA 2022
16 - 19 Dec 22
Chengdu, China

ICIEA22-000387 Text Classification of Judgement Documents Considering Sample Imbalance

Zhaoxu Yang§ , Jike Ge*§, Tingkai Hu§ , Wencheng Yu§ , Yujie Zheng§ , Yan Dong§
Chongqing University of Science and Technology, China

Text Classification of Judgement Documents Considering Sample Imbalance



The overall architecture of our RoBERTa Text-RNN Self-Attention model

In order to get better text representation, we use RoBERTa as the feature encoder to obtain word embedding representation.
Through different masking strategies, the model can learn the information of each key position, thereby constructing a word embedding representation that is more suitable for judgement documents.
After, word representations are obtained, we use bidirectional LSTM (Bi-LSTM) to model long-distance dependencies in the judgement documents, since fact descriptions in judgement documents are generally long texts, and the language is more standard and the article is more logical.
The basic idea of the self-attention mechanism is the model will automatically assign higher weight to important information or word embeddings.

And in order to learn how to classify the hard-to-learn samples, by giving more weight to the loss value of them, the model is more focused on these complex samples, thereby solving the problem of inaccurate classification of categories with few samples. Since there are 202 crime labels and 183 legal labels in the CAIL2018-Small dataset, and the label distribution is extremely unbalanced.
by adding label prior knowledge, which makes the model more inclined to the hard-to-learn samples, speeding up the training to the model.

Summary

In this paper, RTR-SA (RoBERTa Text-RNN model based on the self-attention mechanism) is proposed to cope with the problem of unbalanced distribution of sample categories. First, the proposed method obtains word embedding representations via RoBERTa as the feature encoder, then models long-range dependencies in long texts via Bi-LSTM, and finally models key information using a self-attention mechanism. For the problem of label distribution imbalance, the focal loss function is improved to enhance the model's ability to learn complex samples.