Additional File for SURE

1 ADDITIONAL EXPERIMENT

The framework of SURE can not only be applied to screen ncRNA-protein interaction (NPI) negative samples, but also for other negative sample preparation tasks. In this additional file, a text classification task is experimented with, and excellent performance is achieved. The next subsections introduce the experiment in detail.

1.1 Dataset

We download the IMDb dataset [1]. The IMDb dataset consists of 50,000 IMDb movie reviews, which involve 25,000 positive samples and 25,000 negative samples. Unlike the datasets in NPI, the IMDb dataset contains reliable negative samples. To test the performance of SURE on text classification, we artificially construct unlabeled text datasets. We mix partial positive samples and negative samples to form unlabeled samples. The total number of unlabeled samples is set to 5000. The proportions of positive samples among the unlabeled sample are set to be $0 \sim 0.8$, with each 0.1 as an interval. Therefore, 9 unlabeled datasets with different positive proportions are obtained. We do not make the proportion of positive samples higher than 0.8 because it would be difficult to screen negative samples from unlabeled samples if there are few negative samples in unlabeled samples. We use the 5,000 positive samples that do not include in any unlabeled datasets as the positive dataset. The positive dataset and the 9 unlabeled datasets make up 9 complete datasets.

1.2 Feature Representations

Since the download parameter is selected as 10000-num-words when downloading the IMDb dataset, the dictionary size for all sentences is 10000. We use a 10000-dimensional feature vector and a simple zero-one encoding to represent each sentence. The zero-one encoding is done by this way: first, assign IDs to 10000 words. Then initialize a zero vector of dimension 10000 for each sentence, and add one in the id-th position of the vector when the word is contained in the sentence. Since the feature dimension is different from that in the NPI task, we change the input dimensions of policy network and sample inspector network to adapt the new feature dimension.

1.3 Evaluation Metric

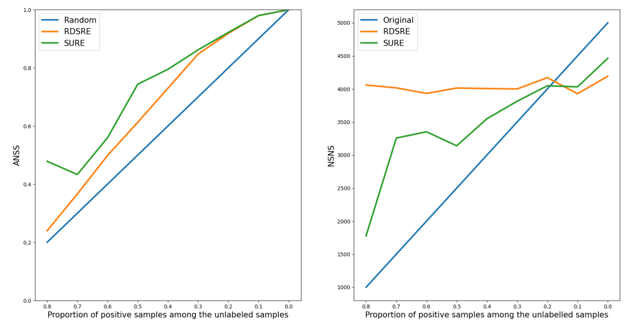
The accuracy of negative samples screened ($ANSS$) can be used as an evaluation metric of the reliability of the screened negative samples, and it is defined in the follows:

$$ANSS = \frac{true\ negative}{screened\ negative} \tag{1}$$

where $true\ negative$ represents the number of screened true negative samples, and $screened\ negative$ represents the number of all screened negative samples.

1.4 Baselines

In text classification tasks, there are currently few negative sample preparation methods. A similar method [2] is able to screen reliable positive samples from the unlabeled samples. This model also bases on deep reinforcement learning. We change this model slightly to enable it to screen reliable negative samples from unlabeled samples and add it as baseline. In this paper, we call this model as RDSRE. We also add random sampling as a baseline.



(a) The ANSS of different methods on IMDb dataset **(b) The NSNS of different methods on IMDb dataset**

Figure 1: Performance of various methods of screening negative samples in terms of the accuracy of negative samples screened (ANSS) and the number of screened negative samples (NSNS) under IMDb dataset. The horizontal coordinate is the proportion of positive samples among the unlabeled samples.

1.5 Performance Comparison

The performance of three methods for screening text negative samples are shown in Figure 1. The observations below are depicted in Figure 3.

- SURE performs better than random sampling and RDSRE. SURE maintains the optimal $ANSS$ for all mixed proportion. Meanwhile, $NSNS$ of SURE is the closest to the number of original negative samples among the unlabeled samples. The lower the positive proportion is, the better the SURE performance.
- As shown in Figure 1, when the proportion of positive samples is less than 0.5, the $ANSS$ of SURE can reach more than 70%. The $ANSS$ is increased by 20% compared with random sampling, and 10% compared with RDSRE. And the lower the positive proportion is, the higher the $ANSS$ of SURE. Therefore, the negative samples screened by SURE are relatively reliable.

From the above results, it can be seen that SURE still has a good performance in screening reliable negative text samples. Therefore, the framework of SURE can be applied to solve negative sample preparation problem for many tasks.

REFERENCES

[1] 2021. IMDb Datasets. https://www.imdb.com/interfaces/. Accessed: 2021-08-30.

[2] Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. arXiv:1805.09927 [cs.CL] Forthcoming.