Task: Sentiment Classification with Deep Learning Technology

1. Introduction

This task aims to evaluate the deep learning techniques for sentiment classification. The data set are collected from both Chinese and English product review web site, and deep learning technology is required to use in the classify process

2. Description of the Task

This task aims to predict the polarity of each review of the data set, and the polarity of each review is binary, either positive or negative.

Data Description: the training data and testing data is xml format, for an instance:

<review id="1">

This drive was a huge disappointment…

</review>

Submission Format: the submission should use the following format:

team-ID: the ID of team

run-tag: the tag of submission, each team could submit 2 result at most.

review-ID: ID of review

polarity: the polarity of review : positive, negative.

For the above example, the submission should be:

team_name 1 1 negative

Evaluation Metric: in the sentiment classification task, metric is based on precision, recall and F1-measure:

TP: The number of the positive samples which are correctly classified.

FP: The number of the positive samples which are falsely classified.

TN: The number of the negative samples which are correctly classified.

FN: The number of the negative samples which are falsely classified.

$$Precision_{pos} = \frac{TP}{TP+FP}, Precision_{neg} = \frac{TN}{TN+FN}$$

$$Recall_{pos} = \frac{TP}{TP+FN}, Recall_{neg} = \frac{TN}{TN+FP}$$

$$F1_{pos} = \frac{2Pression_{pos}Recall_{pos}}{Pression_{pos}+Recall_{pos}}, F1_{neg} = \frac{2Pression_{neg}Recall_{neg}}{Pression_{neg}+Recall_{neg}}$$

3. Data

In this evaluation, both the English data and Chinese data are in xml format, Unicode encoding (utf-16).

4. Evaluation Arrangements

1) The evaluation is off-line conducted.

2) Each participator must utilize deep learning technology in the sentiment classification model.

3) The labeled data is from multiple domains, such as book, DVDs, and electronics. The labels of the samples in the sample data are automatically generated according to the stars given by the users (There are 1-5 stars. Samples with 4 or 5 stars are treated as positive samples while samples with 1 or 2 stars are treated as negative ones). The labels of the samples in the test data are manually checked because a few of them could possibly be incorrectly labeled by the stars.

4) Extra labeled data (either automatically labeled or manually labeled) is not allowed. But extra unlabeled data is allowed to use in the classification model.

5) Standard NLP tools and standard machine learning or deep learning tools, such as word2vec tools, are allowed to use, but sentiment analysis or opinion mining tools and resources (opinion lexicons) are not allowed to use.