

ABSTRACT

SAMPLE CLEANING METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

The present disclosure discloses a sample cleaning method and apparatus, an electronic device, and a storage medium, relating to the technical field of artificial intelligence. The method herein includes: performing semantic clustering on samples in a corpus to obtain sample sets of multiple cluster categories; determining tag categories contained in the sample set of each cluster category, and dividing the sample set of each cluster category into a plurality of sample subsets based on the tag categories; and determining, for the sample set of each cluster category, a mislabeled target sample subset from the plurality of sample subsets, thus implementing a sample cleaning task for the corpus. The technical solution provided by the present disclosure enables rapid localization of mislabeled samples in the corpus, with high accuracy of mislabeled samples determined, thereby effectively reducing time and labor cost investments.

Fig. 1