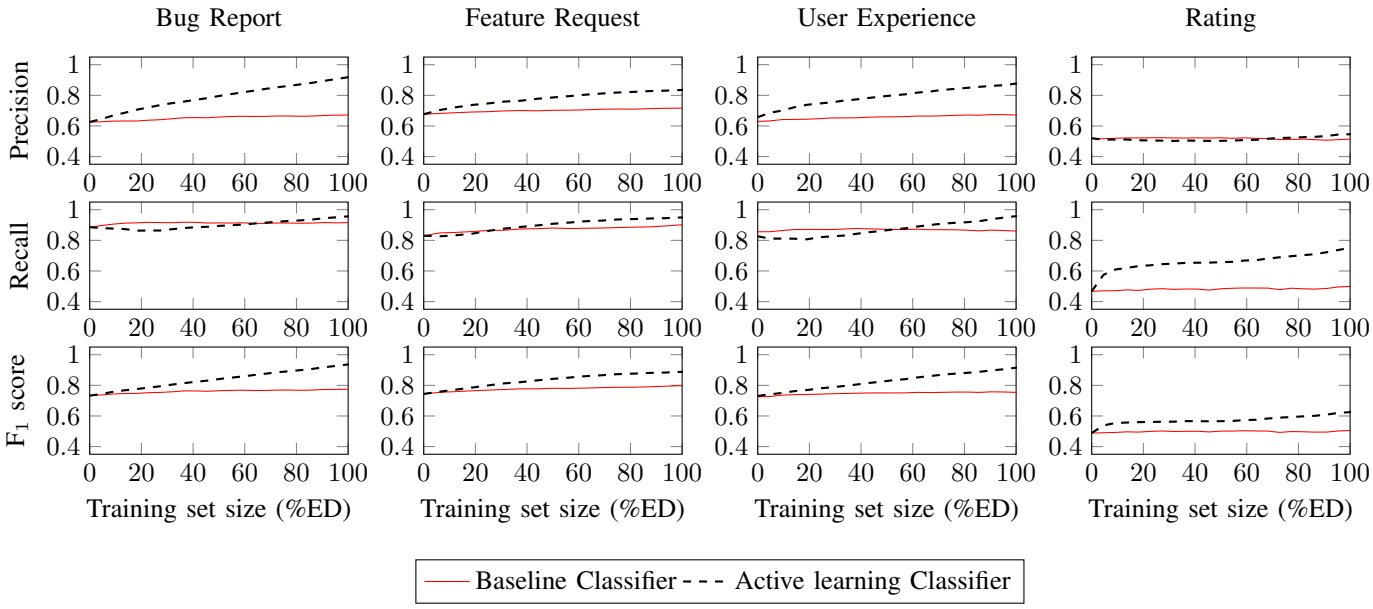

Fig. 4. Comparing the active learning and baseline classifiers in the binary classification task

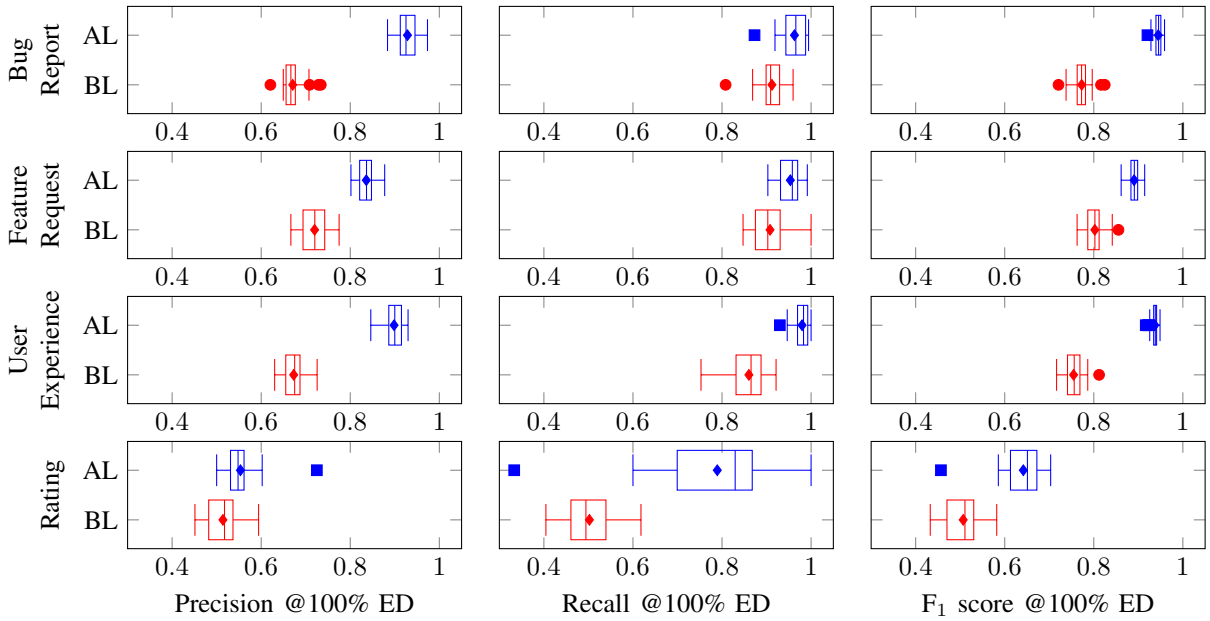

Fig. 5. Comparing the Precision, Recall, and F_1 scores for baseline (BL) and active learning (AL) classifiers for maximum training set (100% ED)

negative instances in the overall dataset, respectively. As a result, when tested, the classifier yields fewer false positives, and thus, higher precision, in each successive iteration.

In contrast, we cannot claim the same pattern for the baseline classifier. Since the baseline classifier adds new training instances randomly, the distributions of reviews in the classes may not be representative of the of entire dataset even when we increase the training set, explaining our observation that precision curves are almost flat for the baseline classifiers.

For the Rating classifier, we note that the Rating class is much larger than other classes (Table III). Further, the class is also "noisy" or less-structured in that there are a variety of ways to merely express a rating. In such cases, we conjecture that active learning, similar to baseline, is unable to pick representative reviews for the training set.

A binary AL classifier yields a higher precision than a BL classifier when the classes are sufficiently well structured. For app review analysis, the Bug Report, Feature Request, and User Experience classes are sufficiently well structured.

3) Recall: For all but Rating classifier, the recall values for both baseline and active learning are high, in general.