Задача: обучить модель классифицировать комментарии на позитивные и негативные. В вашем распоряжении набор данных с разметкой о токсичности правок.

Решение:Мы исследовали размеченный набор текстов и обучили 4 типа моделей определять токсичные комментирии. К сожалению, BERT оказался недоступен. Однако, и без него удалось достичь некоторых результатов.

Самого высокого f1 удалось достичь при помощи логистической регрессии. Выбор лучшего порога классификации помог существенно снизить долю ложноположительных предсказаний, то есть снизить количество несправедливых банов, однако, мы видим, что когда f1 растет за счет снижения ложноположительных предсказаний, число ложноотрицательных становится выше. В такой ситуации рост f1 приводит к снижению строгости, и большее количество токсичных комментаторов имеют возможность продолжать комментировать.