Где можно использовать модель для классификации токсичности сообщений?

Обучающий датасет по большей части состоял из сообщений длины 5-50 слов, так что разумно использовать модель на похожих текстовых данных. Например, можно внедрить модель для классификации токсичных сообщений на форумах или комментариев в социальных сетях. Таким образом перед публикацией новое сообщение прогоняется через модель и крайне токсичные тексты могут не публиковаться в сети, а их автор может получить предупреждение. Также такие сообщения могут отправляться на дополнительную ручную проверку модераторам.

Как стоит внедрить модель в продакшен?

Перед внедрением стоит удостовериться в работоспособности модели и провести АБ-тест. Можно части пользователей оставить предыдущую версию публикации сообщений с некоторой имеющейся логикой, а другой части включить новую модель классификации. Также допустим, что пользователи могут сообщать в поддержку о токсичных по их мнению сообщениях, а также ставить дизлайки на сообщения. Далее можно сравнить количество сообщений с негативными реакциями и количество репортов о токсичных сообщениях и сделать вывод качестве модели.

Как поддерживать модель в продакшене?

Модели машинного обучения со временем ухудшаются, так как они чувствительны к изменениям в реальном мире, нам нужно поддерживать качественную модель.

Во-первых, нужно контролировать, в каком виде поступают входные данные, если размер сообщений будет значительно больше, то скорость обучения снизится, и гарантировать ту же точность будет сложно. Также наша модель изначально обучалась только на сообщениях на английском, поэтому на данном этапе в текстах на другом языке не будут корректно найдены токсичные сообщения, со временем можно создать новую большую модель. Необходимо смотреть и за распределением данных, так как смещения и выбросы также в будущем повлияют на качество модели. Это можно сделать с помощью статистических тестов. Во-вторых, стоит отслеживать крайние случаи, когда модель сильно уверена в ответе, а также добавлять в модель новые данные, полученные с помощью обратной связи пользователей.