Предвзятости и ограничения

Если вы намерены использовать предварительно обученную модель или ее дообученную версию в рабочей среде, имейте в виду, что, хотя эти модели являются мощными инструментами, они имеют ограничения. Самое большое из них заключается в том, что для предварительного обучения на больших объемах данных исследователи часто очищают весь контент, который они могут найти, беря как лучшее, так и худшее из того, что доступно в Интернете.

Для иллюстрации вернемся к примеру пайплайна `fill-mask` с моделью BERT:

```
from transformers import pipeline

unmasker = pipeline("fill-mask", model="bert-base-uncased")
result = unmasker("This man works as a [MASK].")
print([r["token_str"] for r in result])

result = unmasker("This woman works as a [MASK].")
print([r["token_str"] for r in result])
```

```
['lawyer', 'carpenter', 'doctor', 'waiter', 'mechanic']
['nurse', 'waitress', 'teacher', 'maid', 'prostitute']
```

На просьбу вставить пропущенное слово в этих двух предложениях модель дает только один ответ без гендерной принадлежности (официант/официантка). Другие рабочие профессии обычно ассоциируются с одним конкретным полом — и да, проститутка попала в топ-5 вариантов, которые модель ассоциирует с “женщиной” и “работой”. Это происходит даже несмотря на то, что BERT — одна из редких моделей трансформеров, созданная не путем сбора данных со всего Интернета, а с использованием явно нейтральных данных (он обучен на английской Википедии и наборе данных BookCorpus.

Поэтому, когда вы используете эти инструменты, вам нужно помнить, что исходная модель, которую вы используете, может очень легко генерировать сексистский, расистский или гомофобный контент. Дообучение модели на ваших данных не сможет устранить эту внутреннюю предвзятость.