УДК 004.89

ВЫЯВЛЕНИЕ СТРУКТУРНОЙ ПРЕДВЗЯТОСТИ В МОДЕЛЯХ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Голденштейн С., студент направления Computer Visions Master's Degree Programm, II курс
Открытый университет Каталонии, Барселона, Испания

Выявление и устранение структурной предвзятости в моделях обработки естественного языка (NLP) является одной из важнейших проблем, стоящих перед исследователями и разработчиками в области искусственного интеллекта. Структурная предвзятость – это скрытая, систематическая ошибка, возникающая из-за особенностей, заложенных в данных, архитектурах моделей или алгоритмах обучения, которая может приводить к несправедливым или искажённым результатам в процессе обработки текстовой информации. Она может проявляться в различных формах: от несправедливого представления групп и категорий в текстах до искажений в механизмах интерпретации и принятия решений. Такая предвзятость в NLP моделях становится особенно критичной в контекстах, где точность и справедливость решений имеют важное значение, таких как в правовых, медицинских, финансовых и социальных системах [1].

Проблема структурной предвзятости в NLP моделях чаще всего обусловлена не только данными, на которых эти модели обучаются, но и самой архитектурой и дизайном этих систем. Большинство моделей NLP, в том числе широко распространённые трансформеры, такие как BERT, GPT и их производные, обучаются на огромных корпусах текста, которые включают в себя большое количество разнообразной информации, в том числе возможные предвзятости, проявляющиеся в языке, который они используют. Эти предвзятости могут быть как явными, так и скрытыми, и могут приводить к серьёзным социальным и этическим проблемам, таким как усиление стереотипов, дискриминация определённых групп или даже создание ложных ассоциаций, которые могут влиять на выводы моделей.

Один из самых ярких примеров проявления структурной предвзятости в NLP моделях – это генерация стереотипных или предвзятых текстов. Модели, обученные на текстах, наполненных историческими или социальными предвзятостями, могут легко «усваивать» и воспроизводить такие предвзятые отношения, создавая неприемлемые или искажённые интерпретации. Например, модель может автоматически связывать определённые профессии с мужским или

женским полом, расы – с определёнными негативными качествами, а возраст – с определёнными ограничениями. Эти типы предвзятости могут не только уменьшить точность предсказаний модели, но и привести к нежелательным последствиям, когда результаты влияют на реальные социальные процессы, такие как рекрутинг, предоставление кредитов, судебные разбирательства или медицинские рекомендации.

Для выявления структурной предвзятости в NLP моделях применяются разнообразные методы анализа данных и интерпретации моделей. Один из таких методов включает в себя анализ на основе контекстуальной валидации, который позволяет исследовать, как модели реагируют на фразы или предложения, содержащие чувствительные признаки, такие как пол, расу, религию или этническую принадлежность. В ходе таких тестов часто используется параллельная проверка предсказаний, где модель выполняет задачу для разных формулировок, чтобы увидеть, изменяются ли результаты при изменении контекста, связанного с этими признаками. Например, если модель генерирует негативные оценки на основе одного контекста и позитивные на основе другого, то это может свидетельствовать о наличии структурной предвзятости [2].

Ключевым этапом в оценке структурной предвзятости является анализ внимания (attention analysis). В моделях, основанных на трансформерах, например, внимание используется для выделения важнейших частей текста. Исследование того, какие слова или фразы модель считает наиболее важными при принятии решений, позволяет выявить скрытые предвзятости. Например, если модель постоянно «обращает внимание» на слова, связанные с чувствительными признаками (например, пол или этническая принадлежность), это может свидетельствовать о том, что эти признаки играют чрезмерную роль в принятии решений модели, что в свою очередь может создавать структурную предвзятость.

Кроме того, в рамках проверки на предвзятость могут быть использованы методы тестирования генеративных моделей. Например, с помощью контрфактического анализа можно создавать вариации текстов с изменёнными чувствительными признаками (например, заменяя мужской род на женский или наоборот), чтобы проверить, изменяется ли поведение модели. Если предсказания модели значительно меняются в зависимости от изменения таких признаков, это может указывать на то, что модель чрезмерно зависит от этих признаков и может быть предвзятой.

Другим важным методом является использование метрик справедливости в NLP. Для анализа структурной предвзятости разрабатываются специальные метрики, которые позволяют сравнивать предсказания модели по различным демографическим или социальным категориям. Например, может быть проведена проверка, что ошибки модели (ложные срабатывания, ложные

отрицания и т. п.) одинаково распределены между различными группами, независимо от их чувствительных характеристик. Метрики могут также учитывать, насколько модель справедливо генерирует ответы в контексте, связанный с определёнными этническими, социальными или гендерными группами [3].

Для минимизации выявленной структурной предвзятости разрабатываются методы регулирования и обучения с учётом справедливости. Это включает в себя использование методов регуляризации, которые накладывают штрафы на использование предвзятых признаков в процессе обучения модели. Одним из таких методов является обучение с использованием справедливых потерь, где добавляются дополнительные термины в функцию потерь, стимулирующие модель избегать предвзятых решений. Другим методом является использование генерации контрфактических примеров, чтобы обучить модель на данных, которые специально очищены от предвзятости.

Немаловажным аспектом является и постоянный мониторинг предвзятости в уже развернутых системах. В процессе эксплуатации модели могут возникать новые формы предвзятости, связанные с изменением данных, контекста или требований к модели. Для этого необходимо регулярно проверять модель на наличие новых проявлений предвзятости и при необходимости корректировать её поведение [4].

В результате, выявление структурной предвзятости в моделях обработки естественного языка требует комплексного подхода, включающего как анализ данных, так и оценку работы модели через различные метрики справедливости и методы интерпретируемости.

Важно, что такие методы помогают не только снизить риски дискриминации и предвзятости, но и повысить доверие к системам ИИ, обеспечивая их соответствие этическим стандартам и справедливости.

Постоянное внимание к этой проблеме в будущем обеспечит более этичное и справедливое использование технологий искусственного интеллекта в решении реальных проблем общества [5].

Список литературы:

1. Bolukbasi, T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings / T. Bolukbasi, K.-W. Chang, J. Zou [et al.] // Advances in Neural Information Processing Systems. – 2016. – Vol. 29. – P. 4349–4357.

2. Caliskan, A. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases / A. Caliskan, J. J. Bryson, A. Narayanan // Science. – 2017. – Vol. 356, № 6334. – P. 183–186. – DOI 10.1126/science.aal4230.

3. Kurita, K. Measuring Bias in Contextualized Word Representations / K. Kurita, N. Vyas, A. Pareek [et al.] // Proceedings of the First Workshop on Gender Bias

in Natural Language Processing. – Florence : ACL, 2019. – P. 166–172. – DOI 10.18653/v1/W19-3823.

4. Nangia, N. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models / N. Nangia, C. Vania, R. Bhalerao, S. R. Bowman // Proceedings of EMNLP. – Stroudsburg : ACL, 2020. – P. 1953–1967. – DOI 10.18653/v1/2020.emnlp-main.154.

5. Пылов, П. А. Исследовательская модель сильного искусственного интеллекта для решения задачи оптического распознования символов / П. А. Пылов, Р. В. Майтак, А. В. Протодьяконов // Инновации в информационных технологиях, машиностроении и автотранспорте : сб. материалов VI Международной научно-практической конференции, Кемерово, 30 ноября - 01 декабря 2022 года. – Кемерово : Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2022. – С. 265–268. – EDN VBJGWS.