Секция «11.7 Социология, экономика и право СМИ»

Методология измерения алгоритмической предвзятости в российских рекомендательных системах

Научный руководитель – Бодрунова Светлана Сергеевна

Алтухова Валерия Олеговна
Студент (магистр)
Санкт-Петербургский государственный университет, Институт "Высшая школа журналистики и массовых коммуникаций Кафедра периодической печати, Санкт-Петербург, Россия
E-mail: st133191@student.spbu.ru

Рекомендательные системы стали основным каналом медиапотребления в России, однако их работа остается «черным ящиком». Непрозрачность логики алгоритмов создает риски систематических искажений в выдаче (algorithmic bias), ведущих к формированию «эхо-камер» и снижению тематического разнообразия [3]. В российской науке отсутствуют стандартизированные методики количественной оценки такой предвзятости, адаптированные к локальным платформам.

Цель работы – разработать и обосновать методику эмпирического выявления и измерения алгоритмической предвзятости. Задачи исследования: определить операционализируемые критерии оценки; разработать дизайн эксперимента с тестовыми аккаунтами; предложить метрики и процедуру количественного анализа контента, собранного в дневниках наблюдения.

Теоретическую рамку образуют концепция «эхо-камер» К. Санстайна и работы по алгоритмической справедливости [2]. Методологически близки зарубежные эксперименты с тестовыми аккаунтами (sock puppet audits): А. Ханнак и др. [4] изучали персонализацию поисковой выдачи, создавая аккаунты с разными характеристиками; Дж. Кулшрестха и др. [5] анализировали влияние истории поиска на разнообразие новостных рекомендаций. В российской традиции к схожим выводам приходят И.В. Козицин и А.Г. Чхартишвили [1], чье агент-ориентированное моделирование показывает, что рост активности пользователей коррелирует со скоростью формирования эхо-камер, а алгоритмы ранжирования могут блокировать кросс-идеологическую информацию. Для нашей методологии ключевой является возможность оценить, насколько система непредвзята к пользователю, не отдает ли предпочтение одним темам или авторам в ущерб другим.

Мы предлагаем четыре критерия оценки предвзятости, каждый из которых измеряется через набор метрик и фиксируется в двух таблицах наблюдения:

1. Тематическое разнообразие (ТР) показывает широту спектра тем в ленте. Метрика: индекс тематического разнообразия – доля уникальных тем от общего числа материалов. Низкий ТР (<0.2) сигнализирует о формировании «пузыря».

2. Индекс повторяемости источников (ИПИ) отражает степень монополизации ленты ограниченным числом авторов. Метрика: доля материалов от трех самых частотных источников (CR3). Если три канала занимают более 50% ленты, фиксируется высокая концентрация.

3. Скорость формирования эхо-камеры (СФЭ) измеряет, как быстро меняется тематический профиль ленты после целевых действий пользователя. Метрика: количество кликов/лайков на материалы определенной темы, необходимое для того, чтобы доля этой темы превысила 50%.

4. Баланс точек зрения (БТЗ) оценивает представленность разных позиций по социально значимым темам. Метрика: индекс тонального разнообразия – соотношение материалов с позитивной, негативной и нейтральной тональностью (сентимент-анализ).

Стоит разобраться с механикой работы эксперимента на конкретном примере. Рассмотрим его принципы на примере платформы «Дзен» (веб-версия). Ради сбора эмпирических данных создаются три тестовых аккаунта с чистыми куки, имитирующих разные поведенческие стратегии, касающиеся одной поляризирующего смыслового кластера (Например, вопрос об одобрении/осуждении повсеместного использования искусственного интеллекта). Направленность аккаунтов предполагается следующая: «Нейтральный» (клики равномерны и случайны), «За» (клики направлены на одобрение), «Против» (клики направлены на осуждение). Для изоляции используются разные браузеры. Период сбора: 1–14 апреля 2026 г. (исключает праздничные дни, способные влиять на алгоритмы).

Для фиксации данных разработаны две взаимосвязанные таблицы. Таблица 1 – журнал действий пользователя, который фиксирует все действия аккаунта (независимая переменная): ID аккаунта, дата, время, тип действия, тема поста, URL, источник. Таблица 2 – журнал состояния ленты, который фиксирует срезы ленты (зависимая переменная): базовый срез (до действий), после каждой серии действий, ежедневно в 10:00. Фиксируются первые 10 постов с указанием ID, даты и времени среза, номера поста, URL, темы, источника, тональности и признаки спорной темы. Связка двух таблиц по времени и ID аккаунта позволяет рассчитать скорость (СФЭ) как количество действий, после которых доля темы превысила 50%, и силу эхо-камеры (динамику ТР, ИПИ, максимальную долю темы).

Предполагается, что аккаунты с узкой тематикой продемонстрируют быстрое падение ТР и рост ИПИ, что подтвердит гипотезу о формировании «эхо-камер» под воздействием алгоритмической предвзятости. Разработанная методика может стать основой для независимого аудита рекомендательных систем в России и для выработки рекомендаций по повышению цифровой грамотности пользователей.

Источники и литература

1) Козицин И.В., Чхартишвили А.Г. О влиянии активности пользователей онлайновых социальных сетей на формирование эхо-камер // Управление развитием крупномасштабных систем (MLSD'2020): труды Тринадцатой международной конференции. М.: ИПУ РАН, 2020. С. 1874–1882. URL: https://mlsd2020.ipu.ru/proceedings/1874-1882.pdf (дата обращения: 28.02.2026).

2) Deldjoo Y., Jannach D., Bellogin A., Difonzo A., Zanzonelli D. Fairness in Recommender Systems: Research Landscape and Future Directions // User Modeling and User-Adapted Interaction. 2023. Vol. 33. P. 1–50. URL: https://arxiv.org/abs/2205.11127 (дата обращения: 28.02.2026).

3) Flaxman S., Goel S., Rao J.M. Filter Bubbles, Echo Chambers, and Online News Consumption // Public Opinion Quarterly. 2016. Vol. 80, No. S1. P. 298–320. URL: https://5harad.com/papers/bubbles.pdf (дата обращения: 28.02.2026).

4) Hannák A., Sapieżyński P., Kakhki A.M., Krishnamurthy B., Lazer D., Mislove A., Wilson C. Measuring Personalization of Web Search // arXiv preprint. 2017. arXiv:1706.05011. URL: https://arxiv.org/abs/1706.05011 (дата обращения: 28.02.2026).

5) Kulshrestha J., Eslami M., Messias J., Zafar M.B., Ghosh S., Gummadi K.P., Karahalios K. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media // arXiv preprint. 2017. arXiv:1704.01347. URL: https://arxiv.org/abs/1704.01347 (дата обращения: 28.02.2026).