ИИ И СОВРЕМЕННЫЕ МЕТОДЫ ИССЛЕДОВАНИЯ МЕДИАТЕКСТОВ КОНЦА XIX – НАЧАЛА XX ВВ.

Мадина Хабибуллаевна ДАЛИЕВА
DSc, Доцент
Узбекский государственный университет мировых языков

Аннотация. *В статье рассматривается влияние современных методов искусственного интеллекта на исследование медиатекстов конца XIX – начала XX вв. Показано, что цифровые технологии радикально расширяют возможности историко-медиавистического анализа: позволяют работать с массивами оцифрованных источников, восстанавливать повреждённые тексты, проводить автоматическую разметку, семантический поиск и количественное моделирование дискурсивных тенденций эпохи. Особое внимание уделяется использованию методов компьютерной лингвистики (NLP), машинного обучения, нейросетевых моделей для анализа газет, журналов, листовок и других медиаисточников поздней Российской империи и раннего советского периода. Обсуждаются преимущества и ограничения ИИ-подходов, а также перспективы интеграции цифровых гуманитарных методов с традиционными историческими и филологическими практиками.*

Ключевые слова: *искусственный интеллект, медиатексты, цифровые гуманитарные науки, корпусный анализ, компьютерная лингвистика, историческая пресса, оцифровка, машинное обучение.*

Исследование медиатекстов конца XIX – начала XX вв. переживает новый этап благодаря внедрению инструментов искусственного интеллекта (ИИ) и цифровых методов анализа. Печатная пресса данного периода – газеты, журналы, политические листовки, частные издательские бюллетени – долгое время оставалась труднодоступным объектом для системного изучения. Огромные массивы источников были рассредоточены по архивам, хранились в виде хрупких бумажных экземпляров, часто в неполном виде. Ситуация изменилась с появлением широкомасштабных проектов оцифровки и развитием технологий обработки естественного языка, которые сделали возможным анализ исторических медиа в масштабах, невозможных для традиционных филологических подходов [1].

ИИ и этап оцифровки источников

Первым шагом в цифровом исследовании медиатекстов стало использование технологий OCR (оптического распознавания символов). Если ранее эта технология применялась преимущественно для современных текстов, то последние версии систем машинного обучения способны работать с дореформенной орфографией, старославянскими шрифтами, дореволюционными газетными гарнитурами и их многочисленными вариациями [2]. Нейросетевые OCR-модели обучаются на корпусах исторических шрифтов, что существенно повышает точность распознавания и снижает необходимость ручной корректуры.

Особенно значимы такие технологии для изучения массовых газет начала XX века, в которых политические и социальные процессы отражались наиболее динамично. ИИ-модели позволяют автоматически восстанавливать поврежденные участки текста, интерполировать утраченные символы и даже предлагать варианты реконструкции строк по контексту, что приближает исследователей к первозданному виду источника [3].

Корпусный анализ и статистическая реконструкция дискурса

После оцифровки тексты становятся доступными для корпусных методов анализа. Создание специализированных корпусов прессы поздней Российской империи, пажадидской, бухарской и раннесоветской прессы позволяет исследователям изучать частотность лексем, динамику появления тематических групп слов, сетевую структуру цитирования и межтекстовые связи [4].

Методы распределённой семантики (word embeddings) дают возможность выявлять изменения смысловых полей. Например, можно отследить, как менялось значение понятий «прогресс», «народ», «свобода», «реформа» в разные годы и в разных изданиях. Векторные модели фиксируют семантический дрейф и помогают исследователям обнаруживать скрытые идеологические сдвиги, которые не всегда очевидны при традиционном чтении источников [5].

Тематика, идеология и автоматическое моделирование сюжетов

Одной из передовых технологий, внедрённых в гуманитарные исследования, стала тематика-моделирование (topic modeling). Данный метод позволяет выявлять ключевые дискурсивные темы в корпусе газетных статей, определить, какие сюжеты доминировали в определённые годы, какие социальные вопросы находились в центре внимания публицистов. Например, анализ рубежа XIX–XX вв. показывает рост тем о модернизации, политических реформах, образовании, положении женщин и национально-культурной идентичности [6].

ИИ-алгоритмы также дают возможность выявлять идеологические паттерны в медиатекстах – например, различия между либеральными, консервативными и социал-демократическими изданиями. Методы анализа сентимента (sentiment analysis) помогают определить эмоциональную окраску публикаций, выявлять пропагандистские стратегии, тональность политических статей, различия в позициях изданий по реформаторским вопросам [7].

Межтекстовые связи, цитирование и сетевые модели

Современные ИИ-методы позволяют восстанавливать межтекстовые связи в медиапространстве эпохи. С помощью алгоритмов семантического поиска можно обнаруживать скрытые цитаты, аллюзии, повторяющиеся публицистические схемы, перепечатки и ротации материалов между газетами. Создание графов текстовых связей позволяет визуализировать медиасети того времени и анализировать, какие издания играли роль информационных хабов [8].

Так, например, сеть пересечений между региональными газетами Туркестана и столичными изданиями империи показывает интенсивность культурных и политических коммуникаций. Эти связи, ранее скрытые в бумажных архивах, становятся видимыми благодаря цифровым методам.

ИИ и изучение авторских стилей

Технологии стилометрии – области, использующей вычислительный анализ стиля – позволяют исследовать авторство, выявлять особенности идиостиля, дифференцировать индивидуальные и коллективные способы письма. Такие методы особенно полезны для анализа анонимных политических статей, публиковавшихся в период культурных реформ и революции. ИИ-модели определяют характерные паттерны синтаксиса, частотность определённых конструкций, предпочтение отдельных лексических групп [9]. Это расширяет возможности историков и литературоведов при атрибуции текстов и изучении сетевых структур авторов.

Ограничения и риски использования ИИ в историко-медийных исследованиях

Несмотря на очевидные преимущества, использование ИИ имеет ряд ограничений. Во-первых, алгоритмы не обладают историческим контекстом и могут ошибочно интерпретировать семантические сдвиги, если модели обучены на современных данных. Во-вторых, исторические тексты часто имеют плохое качество оцифровки, что приводит к ошибкам распознавания и снижению точности статистических методов.

Кроме того, существует риск «переинтерпретации», когда сложные культурные явления сводятся к статистике. Исследователь должен критически относиться к автоматическим результатам, сопоставлять цифровые находки с историко-филологическими наблюдениями и учитывать культурный фон эпохи [10].

Перспективы развития цифровых медиаисследований

ИИ открывает путь к созданию масштабных мультимодальных архивов: оцифровка газет может дополняться распознаванием иллюстраций, рекламных блоков, верстки. Алгоритмы компьютерного зрения способны классифицировать изображения, обложки, карикатуры и фотографии, которые играли важную роль в медиа той эпохи.

Важным направлением становится создание «умных» корпусов, в которых ИИ будет автоматически связывать тексты с биографиями авторов, историческими событиями, картами, демографическими данными. Это позволит исследователям строить новые типы интерпретаций, объединяющие количественные методы с культурно-историческим анализом.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Underwood, T. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.
2. Springmann, M., & Lüdeling, A. OCR of historical printings with neural networks. *DH Conference Proceedings*, 2017.
3. Smith, R. An overview of the Tesseract OCR engine. *ICDAR*, 2007.
4. McEnery, T., & Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012.
5. Hamilton, W., Leskovec, J., & Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. *ACL*, 2016.
6. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003.

7. Pang, B., & Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2008.

8. Moretti, F. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, 2005.

9. Holmes, D. I. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 1998.

10. Jockers, M. L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.