ИИ-ТЕХНОЛОГИИ В АНАЛИЗЕ ГАЗЕТНЫХ СТАТЕЙ УЗБЕКИСТАНА НАЧАЛА XX ВЕКА

Дилфуза Муминовна ТЕШАБАЕВА
DSc, Профессор
Узбекский государственный университет мировых языков

Аннотация. *Статья посвящена применению современных технологий искусственного интеллекта (ИИ) для изучения газетных текстов, созданных в Узбекистане в первые десятилетия XX века. В центре внимания – возможности автоматической оцифровки, восстановления, разметки и анализа исторической прессы, включая джадидские и раннесоветские издания. Рассматриваются методы компьютерной лингвистики, корпусной аналитики, тематического моделирования, стилометрии и машинного обучения, которые позволяют выявлять дискурсивные особенности эпохи, динамику общественных дебатов, идеологические сдвиги, а также особенности формирования медийного языка. Отмечаются технологические сложности работы с дореформенной графикой, историческими шрифтами и фрагментарными архивами. Подчёркиваются перспективы создания специализированных цифровых корпусов узбекской прессы и интеграции ИИ в междисциплинарные исследования культурных процессов начала XX века.*

Ключевые слова: *искусственный интеллект, историческая пресса, джадидизм, газетные тексты, корпусная лингвистика, стилометрия, машинное обучение, Узбекистан, начало XX века.*

Исследование газетной прессы Узбекистана начала XX века переживает качественно новый этап, связанный с внедрением технологий искусственного интеллекта (ИИ). Джадидские газеты, информационные бюллетени, сатирические журналы, просветительские издания Самарканда, Ташкента, Ферганы и Бухары представляют собой бесценные источники по истории общественной мысли региона. Однако до недавнего времени системная работа с ними была затруднена: коллекции хранились в архивных фондах, тексты были частично утрачены, повреждены, напечатаны нестандартными шрифтами или на разных вариантах турки и чагатайской письменности.

С появлением современных ИИ-инструментов ситуация начала стремительно меняться. Сегодня исследователи получают доступ не только к оцифрованным материалам, но и к возможностям их автоматической реконструкции, анализа и интерпретации в масштабах, которые раньше были невозможны.

ИИ и оцифровка газетных материалов

Технологии OCR на основе нейросетей стали первым значимым этапом в работе с исторической прессой. Классические OCR-системы плохо распознавали дореформенные шрифты, однако модели, обученные на корпусах газет начала XX века, позволяют восстанавливать тексты с высокой точностью [1].

Особенно важны такие технологии для узбекской прессы 1900–1925 гг., где сосуществовали разные алфавитные системы: арабская графика, реформированные варианты письма, ранние попытки латинизации. Нейросетевые модели, адаптированные к историческим типографским гарнитурам, способны не только распознавать буквы, но и исправлять искажённые фрагменты, интерполировать пропуски и повышать качество итогового корпуса [2].

Корпусный анализ и количественные методы исследования прессы

Когда тексты оцифрованы, они становятся доступными для корпусных методов анализа. Создание специализированных корпусов узбекских газет начала XX века открывает путь к изучению словарного состава эпохи, тематических сдвигов и идеологических линий.

Методы частотного анализа позволяют выявлять, какие понятия доминировали в медийном пространстве: *маърифат* (просвещение), *миллат* (нация), *ислохот* (реформа), *жамият* (общество), *илм* (наука). Эти ключевые слова характеризуют интеллектуальную повестку джадидских реформаторов, публицистов и педагогов [3].

Распределённые семантические модели (word embeddings) дают возможность отслеживать изменения смыслового поля отдельных слов. Например, термин *миллат* в ранних джадидских текстах имеет культурно-просветительский оттенок, тогда как ближе к 1920-м годам он получает политизированное содержание. Такие микросдвиги могут быть выявлены автоматически, что является важным вкладом ИИ в историко-лингвистические исследования [4].

Тематическое моделирование: реконструкция дискурсивной повестки

Тематическое моделирование (LDA и его модификации) становится одним из ключевых инструментов анализа массовых источников. Применение этих методов к узбекской прессе начала XX века позволяет выделить устойчивые тематические кластеры: образовательные реформы, критика традиционной школы, новые научные знания, экономические вопросы, вопросы женского образования, социальная модернизация, антирелигиозная полемика раннего советского периода [5].

Каждая тема имеет свою динамику: одни усиливаются к 1917 году, другие исчезают после административных реформ 1920-х гг. Тематическое моделирование помогает количественно описать культурный переход от джадидского модернизационного проекта к идеологии нового государства.

Стилометрический анализ и исследование авторских стилей

Стилометрия – ещё одна важная область использования ИИ-технологий. Она позволяет анализировать идиостиль авторов, сравнивать журналистские школы, выявлять скрытое авторство. Джадидские издания часто публиковали анонимные статьи, особенно в период политических репрессий или острой социальной критики.

Стилометрические модели выявляют повторяющиеся синтаксические конструкции, характерные последовательности слов, предпочтительные морфемные структуры и функциональные слова, что облегчает атрибуцию текстов [6].

Для истории узбекской прессы это особенно актуально, поскольку многие значимые публикации были подписаны псевдонимами или коллективными обозначениями (*муаллиф*, *бир жавон туркестонлик*, *кўзгу* и др.).

ИИ и реконструкция медиапространства эпохи

Важно отметить, что ИИ-методы позволяют выявлять связи между текстами, которые невозможно обнаружить вручную. Алгоритмы семантического поиска показывают, какие статьи перекликались, какие идеи мигрировали между изданиями, как формировались сетевые структуры влияния между Самаркандом, Бухарой, Ташкентом и Ферганой [7].

Такой подход помогает реконструировать коммуникативное пространство эпохи: можно проследить, какие газеты играли роль интеллектуальных центров, какие темы «перекликались» между регионами, какие авторы формировали идеологические линии.

Визуальная аналитика и работа с газетными изображениями

Газеты начала XX века активно использовали визуальные элементы: графику, карикатуры, декоративные виньетки, схемы. Современные методы компьютерного зрения позволяют классифицировать изображения, выявлять повторяющиеся визуальные сюжеты, анализировать карикатурные мотивы и сопоставлять их с текстовой частью [8].

Это открывает возможность комплексного исследования медийного языка эпохи, где текст и изображение дополняли друг друга.

Ограничения и риски цифровых методов

Несмотря на значительный прогресс, необходимо учитывать ограничения ИИ-подходов:

– ошибки OCR при работе с повреждёнными фрагментами,
– трудности нормализации дореформенной лексики,
– риск неверной интерпретации семантических моделей,
– необходимость постоянной валидации результатов традиционными филологическими методами [9].

ИИ способен ускорить анализ, но не заменяет критическое прочтение и историко-культурную интерпретацию.

Перспективы развития направления

Будущее исследований узбекской прессы начала XX века связано с созданием крупных открытых цифровых архивов, мультимодальных корпусов и автоматизированных систем аналитики, интегрирующих текст, изображение, контекст и биографические данные авторов.

Особое значение будет иметь построение «умных карт» интеллектуальных сетей эпохи, позволяющих выявлять траектории идей, связи между реформаторами и соотношение между региональными и общетуркестанскими тенденциями [10].

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Smith, R. An overview of the Tesseract OCR engine. *ICDAR*, 2007.
2. Springmann, M., Lüdeling, A. OCR of historical printings with neural networks. *DH Conference Proceedings*, 2017.

3. McEnery, T., Hardie, A. *Corpus Linguistics*. Cambridge University Press, 2012.
4. Hamilton, W., Leskovec, J., Jurafsky, D. Diachronic word embeddings. *ACL*, 2016.
5. Blei, D. Latent Dirichlet Allocation. *JMLR*, 2003.
6. Holmes, D. Stylometry in historical analysis. *LLC*, 1998.
7. Moretti, F. *Graphs, Maps, Trees*. Verso, 2005.
8. Yanai, K. Image recognition in historical documents. *CVPR Workshops*, 2018.
9. Jockers, M. *Macroanalysis*. University of Illinois Press, 2013.
10. Underwood, T. *Distant Horizons*. Chicago University Press, 2019.