Домбровская Анна Юрьевна, доктор социологических наук, главный научный сотрудник Центра политических исследований, Финансовый университет при Правительстве РФ, г. Москва, Россия; e-mail: aydombrovskaya@fa.ru, ORCID 0000-0002-2015-2349

Бродовская Елена Викторовна, доктор политических наук, главный научный сотрудник Центра политических исследований, Финансовый университет при Правительстве РФ, г. Москва, Россия; e-mail: brodovskaya@inbox.ru, ORCID 0000-0001-5549-8107

ОЦЕНКА РЕЛЕВАНТНОСТИ ИДЕНТИФИКАЦИИ ТЕКСТОВ ОБ ЭМИГРАЦИИ БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ ДЛЯ МОНИТОРИНГА СОЦИАЛЬНЫХ МЕДИА[1]

Актуальность постановки заявленной темы связана с дефицитом методологических экспериментов, фокусирующихся на оценке эвристики искусственного интеллекта в исследовании цифровых маркеров социальных настроений пользователей. В статье анализируются профильные и смежные исследования, измеряющие язык, текст как данные для квалиметрического исследования, особое внимание уделено опыту исследователей, оценивающих репрезентацию цифровых маркеров гражданских установок пользователей социальных медиа. Замысел исследования состоит в оценке релевантности массивов, выгружаемых и сегментируемых нейросетью – языковой моделью LSTM, используемой сервисом «Медиалогия» с помощью специально обоснованных поисковых запросов, и анализе возможных ограничений нейросети в идентификации релевантного контента. В качестве методов сбора и анализа данных обосновываются когнитивное картирование (как предварительный этап исследования для выявления поисковых запросов), социально-медийный анализ с использованием сервиса «Медиалогия», ручной анализ репрезентативной выборочной совокупности выгруженного массива для оценки релевантности автоматически аккумулированного контента. Показаны типичные ошибки в создании поисковых запросов и способы преодоления их в обосновании лингвомаркеров для дальнейшего использования нейросети в ходе выгрузки релевантного датасета. Выявляется соотношение адекватности проведенной языковой моделью идентификации типов контекстов и направленности аккумулированных документов, делается предположение о причинах большей или меньшей релевантности сообщений, выгруженных сервисом.

***Ключевые слова**: социальные медиа, цифровая коммуникация, социально-медийный анализ, когнитивное картирование, нейросети, большие языковые модели, методический эксперимент.*

Как цитировать: *Домбровская А. Ю., Бродовская Е. В.* Оценка релевантности идентификации текстов об эмиграции большими языковыми моделями для мониторинга социальных медиа // Вестник Московского университета. Серия 10. Журналистика. 2025. № 5. С. 209–232. DOI: 10.55959/msu.vestnik.journ.5.2025.209232

Anna Yu. Dombrovskaya, Doctor of Sociology, Chief Researcher, Center for Political Studies, Financial University under the Government of the Russian Federation, Moscow, Russia; e-mail: aydombrovskaya@fa.ru, ORCID 0000-0002-2015-2349

Elena V. Brodovskaya, Doctor of Political Sciences, Chief Researcher, Center for Political Studies, Financial University under the Government of the Russian Federation, Moscow, Russia; e-mail: brodovskaya@inbox.ru, ORCID 0000-0001-5549-8107

EMIGRATION NARRATIVES RELEVANCE ASSESSMENT BY LARGE LANGUAGE MODELS FOR SOCIAL MEDIA MONITORING

The paper's relevance lies in the lack of methodological experiments specifically focused on evaluating the heuristics of artificial intelligence (AI) for assessing users' social sentiments expressed through digital markers, particularly in the context of Russian national audience's emigration intentions. The article delves into studies that measure language and text as data, with a special focus on understanding users' social attitudes and behaviors expressed through digital markers. The primary objective of the study is to assess the significance of the streams downloaded by the neural network, specifically the LSTM language model employed by the Medialogia service. The research design encompasses cognitive mapping (a preliminary stage to identify search queries) and social media analysis conducted by the Medialogia service. Additionally, a representative sample of the downloaded streams is manually analyzed to evaluate their relevance. The paper highlights common errors in creating search queries and provides strategies to overcome these inaccuracies, which can be further utilized to enhance the neural network's ability to download relevant datasets. Furthermore, the relevance of the streams segmentation conducted by the language model is analyzed. The paper makes an assumption about the underlying reasons for the varying degrees of relevance of documents (posts) downloaded by the service.

***Key words**: social media, digital communication, social media analysis, cognitive mapping, neuronetworks, large language models, methodological experiment.*

To cite this article: Dombrovskaya A. Yu., Brodovskaya E. V. (2025) Otsenka relevantnosti identifikatsii emigratsionnykh tekstov bol'shimi yazykovymi modelyami dlya monitoringa sotsial'nykh media [Emigration narratives relevance assessment by large language models for social media monitoring]. *Vestnik Moskovskogo Universiteta. Seriya 10. Zhurnalistika* 5: 209–232. DOI: 10.55959/msu.vestnik.journ.5.2025.209232

DOI: 10.55959/msu.vestnik.journ.5.2025.209232

Введение

В настоящее время человечество переживает масштабный виток технологического развития, связанный прежде всего с повсеместным внедрением технологий ИИ. В связи с этим научное и экспертное сообщества условно разделились на ИИ-оптимистов и ИИ-пессимистов.

В системе аргументов первых такие сильные стороны использования ИИ, как их инклюзивность, благодаря гибридизации цифрового и социального пространств, адаптивность к различным национальным системам, благодаря заложенному принципу машинного обучения, высокая результативность, которая обеспечивается эффектами *Total Data* (Володенков, 2017). Такому прогрессистскому взгляду на интенсивное внедрение технологий ИИ во все сферы жизни общества противопоставлена позиция, согласно которой на первый план выходят различные угрозы и риски, описываемые в широком диапазоне от «антропологического кризиса» (Багдасарян, Володенков, Жмурин, 2025) до десуверенизации национальных государств под натиском транснациональных корпораций или даже до гибели человеческой цивилизации.

Опасения ученых вполне понятны, поскольку сам принцип построения моделей ИИ, эффекты их применения, степень и характер их влияния на общественное мнение задаются природой, наполненностью и принадлежностью баз данных, на основе которых осуществляется машинное обучение. Особенно актуальной эта проблема является на фоне фундаментальной информационной войны между РФ и странами «коллективного Запада», в которой технологии ИИ применяются в качестве когнитивного оружия, направленного на самые различные мишени, среди которых отдельно можно выделить темы-триггеры, вызывающие наибольший резонанс в обществе. Одной из таких тем является эмиграция из страны в силу ее высокого поляризационного потенциала и конфликтогенности в процессе цифровых коммуникаций и крайне негативных социально-экономических и общественно-политических эффектов в целом.

Применение технологий ИИ для исследования данной темы само по себе является вызовом для современной науки, поскольку требует поиска новой методологии, совершенствования методик

и использования неординарных инструментов. Тем более значимым для дата-инженеров, веб-аналитиков, журналистов больших данных является решение задачи оптимизации языковой модели сервисов мониторинга социальных медиа.

В этом контексте одна из наиболее сложных методологических проблем, которую сегодня преодолевают исследователи социально-политических процессов по их цифровым следам, – сфокусированность большинства сервисов для мониторинга социальных медиа, в том числе основанных на возможностях искусственного интеллекта, на потребностях исключительно прикладных областей работы с целевой аудиторией: маркетологии, рекламы, социально-медийного продвижения бренда. Данные сервисы слабо ориентированы на задачи предметного поля политологии, социологии и некоторых других социально-гуманитарных отраслей. Единственным типом проблемного поля социально-политического анализа, в рамках которого эвристика этих инструментов (сервисов для мониторинга социальных медиа) полностью конгруэнтна цели такого анализа – продуктивность цифровой коммуникации политических акторов или лидеров общественного мнения. Работая с названиями или именами данных коммуникаторов, можно получать наиболее адекватные данные в сравнении со сложными случаями социально-политических контекстов (представленность гражданского дискурса, маркеров политического поведения пользователей и т. п.).

Вместе с тем в предметном поле социальных исследований существенно больше тем и вопросов, связанных не с именами собственными, а со специфическими дискурсами и нарративами, оказывающими весомое влияние на социальные установки интернет-пользователей.

Проблема настоящего исследования состоит в оценке точности работы языковой модели *LSTM*, встроенной в сервис для мониторинга социальных медиа, по аккумулированию и структурированию социально-медийных потоков эмиграционного содержания как кейсе, предполагающем работу не с именами собственными в ходе идентификации релевантных документов, а со сложными социальными контекстами.

Обзор смежных исследований

В связи с активным продвижением возможностей нейросетей в социогуманитарных исследованиях в последние годы актуализировалась необходимость проведения методических экспериментов, направленных на оценку точности работы ИИ, языковых моделей по идентификации релевантных документов и определению их содержательных характеристик.

Г. Ле Менс и А. Галлего (Le Mens, Gallego, 2025) представили результаты изучения возможностей и ограничений больших языковых моделей (*GPT-4, Llama 3, MiXtral, Aya*) по идентификации политических взглядов, репрезентированных в цифровом пространстве. Использовалась довольно простая схема: каждой из исследуемых языковых моделей предлагалась команда выявить в предварительно аккумулированном и загруженном в модель массиве политических текстов нарративы левого или правого толка, а также концепты экономической или социальной направленности на 10 национальных языках. Итоги определения типа политического текста были сопоставлены с данными, предоставленными счетчиками – специалистами по ручной обработке текстов. Примечательно, что максимальный уровень точности работы больших языковых моделей достигал 99%, вместе с тем результаты разнились в зависимости от типа авторов текстов и сути задания. К примеру, идентификация концептов социальной политики была несколько более точной, чем экономической. В сравнении с предлагаемым в настоящем исследовании замыслом с использованием аккумулирующей функции сервиса для мониторинга социальных медиа, Г. Ле Менс и А. Галлего совершают предварительное, специально организованное аккумулирование массива интересующих текстов. Сервисы же наподобие «Медиалогии» осуществляют выгрузку релевантных документов автоматически. При этом сервисы решают важную исследовательскую задачу – установление веса, популярности, распространенности данного конкретного типа контента в социальных медиа и цифровых СМИ, что исключено в случае загрузки в нейросеть готового контента для анализа.

Для понимания степени разработанности заявленной исследовательской проблемы важно замечание К. Бенойта (Benoit. 2020), отмечающего сложность социогуманитарных контекстов, редко подвергающихся шкалированию, метрической обработке,

что может тормозить развитие методики социально-медийного анализа для нужд различных областей социального и гуманитарного знания.

В отечественном сегменте релевантных работ наблюдается явный дефицит социальных исследований, основанных на использовании ИИ, в том числе для изучения текстовых массивов, репрезентирующих гражданские позиции пользователей. Методическим фундаментом таких проектов следует считать эмпирические модели, разработанные двумя командами – коллективом ВШЭ под руководством А. С. Ахременко и группой ученых Финансового университета при Правительстве РФ под руководством Е. В. Бродовской. Исследователи из ВШЭ применяют для автоматизированного анализа цифровых маркеров гражданского и политического поведения индикаторы онлайн-сетевых действий (маркеры конкретных поведенческих актов: одобрение контента в формате лайков, эмодзи, придание контенту вирального потенциала в формате репостов и т. д.). Это означает их приверженность фиксации и измерению объективных показателей поведения, репрезентированных в конкретных цифровых актах в отличие от контекстуального, субстанционального анализа, в большей степени зависящего от стратегии интерпретации исследователем содержательных характеристик текста. Для анализа онлайн-сетевых следов социального поведения пользователей фиксация социальных фактов, отраженных в цифровом действии, служит ценной исследовательской перспективой. Коллектив Финуниверситета сосредоточен на содержательных и структурных характеристиках цифровой коммуникации. Цифровые маркеры поведения и взглядов в этом контексте представляют собой показатели масштаба, динамики распространения сообщений, содержащих эти убеждения, каналы их продвижения, характеристики их целевой аудитории и, что наиболее важно, – семантическое ядро, семантическую структуру релевантных цифровых потоков (Бродовская, Домбровская, Пырма, Азаров, 2020).

Возможности парсера поисковых запросов *GoogleTrends* в построении социальных прогнозов, в том числе эмиграционного процесса, изучал Г. Т. Броницкий (Bronitsky, 2024). Г. И. Глущенко анализировал феномен виртуальной миграции (2021). Набирающее популярность явление цифрового кочевничества исследовали Ю. Г. Волков и соавторы (2021). Маркеры миграционных процес-

сов, отображенных в социальных онлайн-сетях анализировались в работах А. А. Гребенюка и А. А. Субботина (2021), а в трудах О. Искусных акцент делался на глобальной сети как факторе продвижения эмиграционных текстов (2021).

Для нашего исследования труды об удаленной работе важны с позиций разработки словаря лингвомаркеров, учитывающего различные типы эмиграции, в том числе цифровое кочевничество (Aroles, Bonneau, Bhankaraully, 2022; Cook, 2023).

Отдельно следует упомянуть работы Н. Вельша и А. Рафтери, нацеленные на разработку моделей изучения медианарративов о возможностях миграции (Welch, Raftery, 2022). Ключевым для настоящего исследования служит исследование Т. Пуи и соавторов (Pu, Huang, Zhang, Yang et al., 2024), отражающее возможности искусственного интеллекта в сценировании и прогнозировании миграционных процессов и показывающее потенциал нейронных сетей в изучении проблем международной миграции.

В целом следует говорить о довольно немногочисленном пуле зарубежных исследований, применяющих нейросети, элементы ИИ для изучения эмиграционных медианарративов. Специальный анализ возможностей ИИ в аккумулировании информационных потоков эмиграционного содержания, их сегментировании и выявлении семантического ядра, по сути, не предпринимался, вместе с тем перечисленные труды служат прочной основой для разработки методики исследования потенциала ИИ в цифровом измерении эмиграционных нарративов в социальных медиа.

Методика исследования

Методической основой исследования служит *Predictor Mining* (Бродовская, Домбровская, Пырма, Азаров, 2020) – подход, который в общественных науках понимается как интеллектуальный поиск цифровых предикторов социальных и политических процессов. Этот подход подразумевает использование специального программного обеспечения, аккумулирующего по заранее обоснованным лингвомаркерам информационные потоки социальных медиа, релевантные предметному полю исследования.

Замысел настоящего методического эксперимента состоит в проверке результатов идентификации и анализа социально-медийных документов, репрезентирующих эмиграционные нарра-

тивы, выполненных языковой моделью, встроенной в сервис для мониторинга социальных медиа «Медиалогия» посредством репрезентативной выборочной ручной обработки содержательных характеристик полученного датасета. Ручной анализ выгрузки, таким образом, покажет степень чистоты и корректности данных, предоставляемых теми опциями «Медиалогии», в которых используются элементы ИИ. Контроль релевантности и адекватности результатов работы сервиса осуществлялся по двум ключевым параметрам – контекста и направленности (модальности) исследуемых документов. Направленность представляет собой тональность сообщения в отношении эмиграции и принимает следующие значения: позитивная направленность (нарратив, содержащийся в документе, отражает положительные смыслы в отношении эмиграционных установок граждан), нейтральная (нарратив, содержащийся в документе, не позволяет отнести его ни к положительным, ни к негативным смыслам в отношении эмиграционных установок граждан), негативная (нарратив, содержащийся в документе, отражает негативные смыслы в отношении эмиграционных установок граждан). Анализ контекстов представляет собой комплексный параметр, предполагающий установление степени чистоты (релевантности) аккумулированных сервисом документов по трем сегментам: цель, геолокация эмиграции, конверсия эмиграционных интенций.

Отдельно следует отметить то, каким образом осуществлялось обоснование данных сегментов и лингвомаркеров для их автоматического аккумулирования.

Авторский коллектив настоящего исследования реализует следующий порядок действий для обоснования необходимых лингвомаркеров:

1) Поиск и структурирование наиболее влиятельных социально-медийных ресурсов, содержащих маркеры исследовательского предмета.

2) Когнитивное картирование сообщений данных цифровых ресурсов, представляющее собой выявление корреляций между значимыми элементами изучаемых текстовых массивов и установление контекстов, составляющих смысловое поле анализируемого предмета исследования.

3) Обеспечение каждого выявленного контекста соответствующими лингвомаркерами.

4) Тестирование лингвомаркеров с применением сервиса для мониторинга социальных медиа.

5) Уточнение лингвомаркеров для каждого контекста в ходе нескольких итераций повторных выгрузок массивов информации из социальных медиа с целью получения релевантного и большого объема сообщений, содержащих искомые цифровые следы.

Для предварительного этапа полевого исследования – когнитивного картирования релевантных документов – была сформирована выборочная совокупность пабликов, сообществ, каналов двух наиболее популярных социально-медийных платформ РФ: *Telegram* и «ВКонтакте».

Для каждой платформы установлен перечень из релевантных предмету исследования наиболее популярных пабликов/сообществ/каналов. Всего групп – исследовательских кейсов – 250. После первичного анализа контента этих сообществ отобрано 60 сообществ, являющихся наиболее влиятельными (наибольшее число подписчиков) и релевантными предметному полю исследования.

Отбор конкретных документов для внесения в базу данных осуществлялся с использованием целевой выборки и с учетом хронологической репрезентации материалов: глубина отбора 2 года (январь 2023 – март 2025 г.), целевому отбору подлежали документы, датированные срединной неделей (по 2–3 документа каждого дня этой срединной недели) срединного месяца каждого квартала; критерии релевантности: наличие возможности идентифицировать характеристики документа в соответствии с матрицей когнитивного картирования (содержательные, включающие понятные смыслы и согласующиеся с темой исследования материалы). В каждом сообществе/канале – исследовательском кейсе – отбору подлежат 10 материалов (по 2 в первом и третьем кварталах, по 3 – во втором и четвертом кварталах), общее число проанализированных документов – 600.

Исходным фильтром поиска релевантных групп было словоупотребление «за рубеж», позволившее инициировать процесс идентификации подходящих цифровых ресурсов и составить перечень слов – фильтров для дальнейшего поиска нужных сообществ, блогов, пабликов. На данном этапе весьма полезными оказываются возможности рекомендательных систем социально-медийных платформ, существенно расширяющих круг поиска

и идентификации релевантных ресурсов. Отбору подлежат только те социально-медийные сообщества, паблики, блоги, которые отвечают требованиям соответствия содержания предмету анализа, интенсивности и регулярности публикационной активности (отбираемый источник должен поддерживать частоту опубликования материалов не менее 1 документа в неделю вплоть до момента исследовательского среза), существенного охвата аудитории (отбору подлежали социально-медийные ресурсы, располагающие не менее чем 10 000 подписчиков).

База данных когнитивного картирования обработана с помощью программы *IBM SPSS Statistics 23.0*. Кластеризация базы данных, осуществленная с помощью метода *k*-средних, позволила обосновать сегментирование информационного потока об эмиграции в три раздела по типу репрезентированного в них контекста. Основания кластеризации: наличие в документе маркеров обсуждения пользователями работы или учебы за рубежом в контексте временного переезда за границу (наличие/отсутствие); наличие в документе маркеров нарративов о возможности переезда пользователей за рубеж на постоянное место жительства (наличие/отсутствие); наличие в документе указания на топонимы как направления эмиграции (наличие/отсутствие); наличие в документе индикаторов совершенных (планируемых к осуществлению) конкретных действий пользователей по организации переезда за границу. Выявленные кластеры документов показали наличие трех ключевых контекстов, вокруг которых строятся потоки об эмиграции в социальных сетях: цель (временный переезд или выезд на постоянное место жительства), геолокация эмиграции (конкретные направления релокации); конверсия эмиграционных интенций (наличие и отсутствие поведенческих маркеров организации пользователями своего переезда за рубеж).

На основе анализа документов, принадлежащих к каждому отдельному контексту, были идентифицированы наиболее часто используемые авторами сообщений слова и словосочетания, присутствие которых в документе означало релевантность той или иной цели геолокации и наличие конверсии эмиграционных интенций пользователей социальных сетей[2].

В категории «цель эмиграции», помимо переезда на постоянное место жительства и временного пребывания за рубежом с целью учебы или работы, отдельно был учтен феномен «цифрового кочевничества», который предполагает работу на зарубежное госу-

дарство, однако может и не зависеть от места пребывания в течение такой работы. Для выгрузки сегментов массива «цель эмиграции» были использованы соответствующие поисковые запросы.

Значениями категории «геолокация эмиграции» служили названия государств – релокационных дестинаций. В числе кодов этой шкалы – страны Европы, Северной Америки, Латинской Америки и Азии, являющиеся наиболее популярными направлениями эмиграции согласно данным предварительного этапа ручного анализа релевантного контента социальных медиа. Сегмент «геолокация эмиграции» содержал соответствующие топонимы, а также все возможные словосочетания о переезде в различные страны.

Категория «конверсия эмиграционных установок» измерялась бинарной шкалой, значениями которой служили наличие или отсутствие индикаторов конверсии эмиграционных установок, то есть указания на то, выражены ли в нарративе исследуемых документов маркеры конкретных действий граждан по организации своего переезда в другую страну. Для сегментов «отсутствие конверсии эмиграционных интенций» и «наличие конверсии эмиграционных интенций» использованы релевантные поисковые запросы.

Особо подчеркнем, что для поисковых строк каждого контекста применялась опция учета конкретного контекста, предоставляемая используемым сервисом.

Произведенный с помощью поисковых запросов мониторинг социальных медиа (инструмент – «Медиалогия») осуществлен за период с 01.04.2024 по 31.03.2025 гг., аккумулировано 255 249 документов суммарно по всем указанным сегментам с 53 социально-медийных площадок[3], наиболее популярных среди российских пользователей.

Выгруженные массивы по каждому типу контекстов рассматривались как отдельные генеральные совокупности, для оценки релевантности и качества которых применялась систематическая выборка. Размер выборочной совокупности (n) рассчитывался по формуле:

$$\frac{N}{(0{,}03)^2 N + 1},$$

где N – размер генеральной совокупности, 0,03 – относительная погрешность выборки.

После вычисления размера выборки для каждого отдельного массива выгруженных текстов с конкретными типами контекстов рассчитывался шаг отбора для применения систематического выборочного метода (шаг отбора есть частное от деления размера генеральной совокупности на число выборочной совокупности). В таблице 1 показан размер выборочной совокупности для массивов каждого конкретного типа контекста. Общее число проанализированных вручную документов в ходе оценки релевантности выгруженного сервисом «Медиалогия» контента составило 5337 сообщений.

Результаты исследования и их обсуждение

Для понимания и получения возможности интерпретации достигнутых результатов вначале важно проанализировать принцип работы модели нейросети, использующей слои долгой краткосрочной памяти (*LSTM*), заложенной в функционал сервиса «Медиалогия». *LSTM* – дословно *longshort-term memory* (долгая краткосрочная память) – служит версией рекуррентной нейронной сети. Преимущество этого вида нейросети – способность обучаться долгосрочным зависимостям, то есть их практически не нужно обучать (Hochreiter, Schmidhuber, 1997). Такая способность нейросети этого типа критически важна для аккумулирования документов из социальных медиа по созданным запросам, поскольку это означает возможность выгрузить релевантные предмету исследовательского интереса информационные потоки, соответствующие заданному контексту. Входные условия для реализации этой возможности 1) максимально полные, корректные и верифицированные лингвомаркеры – слова и словосочетания, использование которых в документе означает релевантность этого источника информации анализируемому предмету исследования; 2) правильное оформление запросного алгоритма в окне «тематического отчета». Создатели «Медиалогии» убеждают нас в том, что чем более детальную информацию исследователь соберет об интересующем информационном потоке, тем более релевантный массив данных будет результатом автоматической выгрузки.

В таблице 1 показаны результаты оценки релевантности выгруженных «Медиалогией» массивов социально-медийных документов по различным контекстам и подконтекстам эмиграционных интенций (погрешность выборки $\delta = 0{,}03$).

Таблица 1

Оценка релевантности выгруженных «Медиалогией» массивов социально-медийных документов эмиграционных интенций

Контексты и подконтексты	**Число аккумулированных документов в разделе (N) Размер выборочной совокупности (n) Процент релеватных документов (а)**
Контекст «Цель эмиграции»	
«Временный переезд или цифровое кочевничество»	N = 104 363 n = 1 099 а = 91,1%
«Переезд за рубеж на постоянное место жительства»	N = 11 596 n = 1 054 а = 89,3%
Итого: общее число выгруженного массива по контексту «Цель эмиграции» (Nц), совокупной выборки по этому контексту (nц) и средний процент релевантности по всем подконтекстам (ац ср)	Nц = 115 959 Nц = 2 153 ац ср = 90,2%
Контекст «Геолокация эмиграции»	
«Геолокация эмиграции, Европа Северная Америка. Латинская Америка. Азия»	N = 27 730 n = 1 067 а = 93,8%
Контекст «Конверсия эмиграционных интенций»	
«Наличие конверсии эмиграционных интенций»	N = 20 414 n = 1 019 а = 90,8%
«Отсутствие конверсии эмиграционных интенций»	N = 91 146 n = 1 098 а = 94,7%
Итого: общее число выгруженного массива по контексту «Конверсия эмиграционных интенций» (Nк), совокупной выборки по этому контексту (nк) и средний процент релевантности по всем подконтекстам (ак ср)	Nк = 111 560 nк = 2 117 ак ср = 92,8%
Итого: общее число выгруженного массива (Nобщ), совокупной выборки (nобщ) и средний процент релевантности по всем документам разной направленности (аср)	Nобщ = 255 249 nобщ = 5 337 аср = 92,3%

Источник: составлено авторами на основе собственных расчетов и данных сервиса «Медиалогия».

Результаты анализа уровня релевантности документов, извлеченных сервисом по разным контекстам, показывают несколько более высокий уровень соответствия предмету анализа сообщений, относящихся к контексту геолокации эмиграции. Это объясняется изначальной направленностью сервисов для мониторинга социальных медиа, встроенной в них языковой модели на работу с так называемыми поисковыми запросами – брендами. Под брендом подразумевается любое имя собственное: название организации (события), имя человека, топоним. В этой связи используемые в поисковой строке конкретные названия стран и регионов мира позволяют достичь высокой степени релевантности аккумулируемого массива. Вместе с тем задача затрудняется тем, что выгружаемые сообщения должны содержать эти топонимы – лингвомаркеры в нужном контексте. Разработчики сервиса «Медиалогия» заявляют о способности встроенной в сервис языковой модели *LSTM* учесть необходимый контекст закладываемых в поисковую строку топонимов. В данном случае в алгоритме поискового запроса был использован контекст эмиграции и релокационных интенций. Все же наличие нерелевантных документов в полученном массиве свидетельствует о некоторых неточностях в работе нейросети данного типа.

Ручной анализ автоматически аккумулированного контента в разделе «Геолокация эмиграции» выявил нарративы, ошибочно выгруженные сервисом и не являющиеся релевантными для данного конкретного контекста. Примеры неточности в работе нейросети показаны в таких фрагментах, установленных как нерелевантные документы:

«Вы свалили в Испанию, вот там предлагайте инициативы». «Очень верное решение. Ещё надо закон о выдаче всех, кто свалил в Европу». «Знакомые, которые <...> свалили в Германию, регулярно жалуются, что все дорого».

Данные типы документов могут относиться скорее к контексту конверсии эмиграционных интенций, поскольку в них присутствуют нарративы обсуждения возможных эффектов, последствий эмиграции и отношения к релокантам.

Показательным также является сопоставление результатов наиболее показательной опции сервиса для понимания доминирующих геонаправлений эмиграции – облака тегов – и наиболее часто употребляемых слов в выгруженном контенте, выявленных

в ходе предварительно проведенного когнитивного картирования релевантных социально-медийных текстов. В массивах, выгруженных нейросетью на основе лингвомаркеров – топонимов, указывающих на интересующую пользователей дестинацию, явно преобладают те же топонимы, что и выявленные в результате ручного анализа: «Европа» и конкретные европейские страны (21% от числа упоминаний, согласно опции облака тегов «Медиалогии» и 26% – согласно итогам когнитивного картирования); «Северная Америка», «США» и «Канада» составили 8% от числа упоминаний, согласно опции облака тегов «Медиалогии» и 11% – согласно итогам когнитивного картирования). При этом отметим, что около двух третей релевантных документов не содержали вовсе никаких упоминаний топонимов. Можно признать, что различия в долевой представленности указанных топонимов, полученных в ходе ручного и автоматизированного анализа, не столь существенны, как минимум установлены общие для этих расчетов наиболее представленные значения: «отсутствие упоминания топонимов», «названия стран Европы», «названия стран Северной Америки».

Довольно высокую степень релевантности установленному контексту продемонстрировал массив документов, относящийся к разделу «Отсутствие конверсии эмиграционных интенций» (см. табл. 1). Специфика этого контекста состоит в гораздо большей смысловой широте, включающей многочисленные концепты, вплоть до простого упоминания эмиграционной повестки, репрезентации отношения к релокации и самим эмигрантам, опубликования любой информации о возможностях эмиграции и нек. др. В этой связи данный массив – один из самых весомых в численном выражении разделов – более 100 000 выгруженных документов (см. табл. 1).

Довольно сложный тип контекста, предложенный языковой моделью *LSTM* для идентификации и выгрузки, – контекст «Наличие конверсии эмиграционных интенций». В этот раздел должны были быть аккумулированы сообщения, содержащие маркеры конкретных действий по организации переезда за рубеж или уже совершенного акта эмиграции. Предварительный качественный анализ языка, конкретнее – семантического ядра языка, на котором происходит выражение данных поведенческих установок или совершенных действий, показал, что упоминание различных ти-

пов виз и конкретных механизмов выезда за рубеж на временное или постоянное проживание означает выражение конверсии эмиграционных интенций, готовности или наличие опыта релокации. Вместе с тем нейросеть *LSTM* пока допускает некоторые неточности в извлечении контента, выгружая как релевантные документы, содержащие, к примеру, словосочетания *passeport talent,* «паспорт таланта», *protezione speciale, non lucrativa, digital nomad visa, heartland visa, green card usa,* «программа культурного обмена *Au Pair*», «золотая виза», «европейская виза», но контекстуально не соответствующие разделу «Наличие конверсии эмиграционных интенций». Примеры таких неточностей: «Трамп обещал золотую визу ... за 5 мил. долларов». «В США за день продали 1000 золотых виз». «А вы знаете, что сегодня у шенгенской визы юбилей: ей исполняется 30 лет».

Все же, несмотря на эти отдельные ошибки идентификации, «Медиалогия» дает довольно высокий процент релевантности этого сложного контекста для автоматической выгрузки из социальных медиа (см. табл. 1).

Раздел «Цель эмиграции» в целом также довольно чистый, однако контекст «Переезд на постоянное место жительства» имеет наименьший процент релевантности среди всех аккумулируемых контекстов (см. табл. 1). Этот контекст потребовал особой обработки в ходе подготовки алгоритма поисковой строки. Использование словоупотребления «ПМЖ» (постоянное место жительства как цель переезда) привело к выгрузке большого массива нерелевантных документов, в которых «ПМЖ» используется как специальный термин на языке любителей домашних животных, означающий просьбу приютить животное. Также проблематичным для нейросети оказалось выгрузить документы, в которых присутствует маркер «ВНЖ», в контексте эмиграции.. Примером таких нерелевантных сообщений являются посты о получении иностранцами вида на жительство в РФ. Применение в ходе алгоритмизации поисковой строки возможностей исключения из выгрузки документов с нерелевантным контекстом позволило очистить аккумулированный массив сообщений, достичь большей чистоты датасета. Другим сложным типом лингвомаркеров оказались слова и словосочетания, рефлексирующие замужество/создание семьи как цель эмиграции. Потенциальные поисковые запросы данного

элемента контекста были сопряжены с выгрузкой такого типа контента, который содержался в цифровых ресурсах, находящихся за гранью легитимности. В этой связи получение точных структурных данных о соотношении между различными целями эмиграции в выгруженном массиве представляется затруднительным, вместе с тем выведение за скобки этого контекста позволяет сопоставить вес и пользовательский резонанс в отношении потоков разных контекстов и подконтекстов.

Несмотря на эти и некоторые другие затруднения автоматического аккумулирования, нейросеть *LSTM* убедительно показала свои возможности в отношении выгрузки документа, являющегося релевантным контексту «цифровое кочевничество» (как элемента контекста цели эмиграции), но не содержащего этого словосочетания: «*Digital Nomad Visa*: работать из любой точки мира, жить в Дубае».

Это наиболее показательный пример возможностей языковой модели долгой краткосрочной памяти, которая «помнит контекст» и может распознать и идентифицировать как релевантный контент, не содержащий «ключ» (поисковые слова), но контекстуально соответствующий запросу.

Заслуживает интереса также сопоставление доли весов массивов документов, в которых отражены контексты «Наличие конверсии эмиграционных интенций» и «Отсутствие конверсии эмиграционных интенций», полученные посредством ручного анализа в ходе предварительного когнитивного картирования текстов социальных медиа, нацеленных на обсуждение эмиграционной тематики и предоставленных сервисом «Медиалогия». Так, согласно данным таблицы 1, из 111 560 документов, идентифицированных и выгруженных нейросетью как релевантные контексту «Конверсия эмиграционных интенций», 20 414 (18,3%) принадлежат разделу «Наличие конверсии эмиграционных интенций» и 91 146 (81,7%) относятся к разделу «Отсутствие конверсии эмиграционных интенций». Ручной анализ также показал доминирование сообщений, в которых отсутствуют маркеры конверсии эмиграционных интенций. Вместе с тем когнитивное картирование выявило несколько иное соотношение между эти видами контекстов. На рисунке 1 видно, что пропорция между потоками с конверсией и без нее примерно 1 к 3 (37,7% и 62,3%, соответственно), то есть, согласно ручной качественной обработ-

ке, данное соотношение чуть более смещено в сторону представленности в анализируемых текстах социальных медиа конверсии эмиграционных интенций.

В таблице 2 представлены результаты оценки степени релевантности направленности (модальности, тональности) выгруженных сервисом «Медиалогия» документов.

Таблица 2

Оценка релевантности идентификации тональности выгруженных «Медиалогией» массивов социально-медийных документов эмиграционных интенций

Тональность нарратива документов	**Генеральная совокупность по конкретному типу документов (N) Размер выборки для данного типа документов (n) Процент релевантности идентификации направленности нарративов документов (a)**
Позитивная тональность	N = 56 155 n = 1 174 a = 98,3%
Нейтральная тональность	N = 116 904 n = 2 444 a= 99,3%
Негативная тональность	N = 82 190 n = 1 719 a = 95,3%
Итого: общее число выгруженного массива (Nобщ), совокупной выборки (nобщ) и средний процент релевантности по всем документам разной тональности (aср)	Nобщ = 255 249 nобщ = 5 337 aср = 97,6%

Как показывает опыт исследователей, применяющих автоматизированный социально-медийный анализ с помощью языковых моделей сервисов для мониторинга социальных медиа, такая опция, как тональность (модальность, направленность), наиболее востребована в случае работы с уже ранее упомянутым типом поис-

ковых запросов – брендами (именами собственными). В этом случае в качестве результатов соотношения позитива, нейтральности и негатива выступает, по сути, долевое распределение между степенью лояльности цифровой аудитории в отношении того или иного бренда. В случае таких сложных контекстов, как специфика гражданского, политического поведения, в том числе репрезентации эмиграционных установок, опция тональности определяет отношение к изучаемым контекстам. Другими словами, документы, идентифицированные как имеющие позитивную тональность, отражают положительное отношение их авторов к эмиграции и эмиграционным интенциям, а отнесенные к числу контента с негативной модальностью – отрицательное отношение их авторов к данному социальному явлению.

Отметим в этой связи, что в случае работы со сложными социально-политическими контекстами нейросеть позволяет получить информацию о структуре релевантных информационных потоков с точки зрения соотношения различных типов документов массива, в частности долевого распределения типа направленности сообщения.

Сопоставление результатов нейросети и итогов ручного определения тональности говорит о довольно высоком уровне релевантности документов, автоматически идентифицированных как имеющие позитивную, нейтральную или негативную направленность. Довольно немногочисленные ошибки языковой модели *LSTM* вполне укладываются в пределы относительной погрешности не более 3%.

Иллюстративным примером корректной идентификации нейросетью документов с негативной тональностью, имеющих сложный смысловой контекст, служат такие сообщения:

«Выбрав эмиграцию, проиграла. Релоканты-иноагенты захотели на Родину».

«А не пора ли .... указать им свое место, чтобы они ..из России ... были холопами запада».

Подводя промежуточные итоги эксперимента, нацеленного на тестирование возможностей автоматического, с применением нейросети сервиса «Медиалогия» аккумулирования, определения тональности и кластеризации релевантных предмету исследования документов социльных медиа, перечислим следующие выводы. Внедрение в функционал сервисов для мониторинга

социальных медиа элементов ИИ существенно улучшило процессы сбора и анализа информационных потоков, способствовало получению более чистой, релевантной выгрузки, однако проведенный эксперимент подтверждает необходимость сочетания возможностей нейросети с основательной и тщательной работой по восприятию и интерпретации аккумулируемых данных интеллектом исследователя. На сегодняшний день качество информации, получаемой с помощью сервисов для мониторинга социальных медиа, зависит от применения адекватной методики обоснования поисковых запросов, в первую очередь для выгрузки информационных потоков со сложными контекстами, предполагающими работу не с названиями брендов, а с лингвомаркерами различных дискурсов и нарративов. Важным аспектом использования возможностей нейросети в ходе аккумулирования релевантного датасета служит ручной анализ выгружаемого контента с целью выявления поисковых запросов, которые провоцируют попадание в выгрузку документов с нерелевантным контекстоми, принятие соответствующих мер по уточнению алгоритма поисковой строки.

Выводы

Внедрение элементов ИИ в работу сервисов для мониторинга социальных медиа, как показал кейс «Медиалогии», существенно усовершенствовало процессы аккумулирования, определения тональности и кластеризации массивов выгружаемых документов. Если до инкорпорирования возможностей нейросети в функционал сервиса лишь не более 5% аккумулируемого потока сообщений идентифицировалось по типу тональности, при этом порядка 95% документов относились системой к числу нейтральных, что в ходе ручной проверки, разумеется, представлялось поверхностным и неточным, то после внедрения элементов ИИ абсолютное большинство сообщений выгрузки (до 98–99%) идентифицируются как имеющие позитивную или негативную тональность. Для социально-политических исследований, не работающих с названиями брендов (для выявления эффективности цифровой коммуникации политических, коммерческих, общественно-гражданских акторов), как в случае с цифровыми следами эмиграционных установок молодежи, такой прорыв сервисов в установлении тональ-

ности контента социальных медиа играет значимую роль, поскольку выявляет соотношение между эмиграционными и антиэмиграционными, мобилизационными и демобилизационными, мотивирующими и демотивирующими нарративами в информационных потоках релокационной направленности.

Вместе с тем на текущий момент использование сервисов для мониторинга социальных медиа требует тщательной, детальной проработки поисковых запросов и их алгоритмизации в поисковом окне. Кроме того, несмотря на существенные позитивные подвижки, связанные с учетом контекстов сообщений, идентифицируемых «долгой кратковременной памятью» нейросети *LSTM* как релевантные для конкретной выгрузки, все еще актуальна необходимость осуществления нескольких итераций для уточнения «неработающих» лингвомаркеров или коррекции алгоритма поисковой строки для исключения из аккумулируемых данных «ненужных контекстов», внесения прочих требуемых изменений.

Несомненно, новые аналитические возможности данных сервисов, обусловленные внедрением элементов ИИ, повышают научную и практическую ценность результатов социально-медийного анализа цифровых следов социально-политических явлений и процессов. При условии учета всех необходимых требований к использованию такого рода инструментов социально-гуманитарное знание приобретает в лице сервиса со встроенной нейросетью важные средство социальной диагностики.

Примечания

[1] Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финансовому университету при Правительстве РФ.

[2] Отметим, что определение контекстов/категорий, переменных и значений измеряемых признаков осуществляется триангуляционной (в данном случае междисциплинарной) группой экспертов (политологов, социологов, психосемантиков), оценивающих также и адекватность поисковых запросов для аккумулирования потоков по типам контекстов.

[3] Площадки, из которых осуществлялось аккумулирование (выборочно): «ВКонтакте», «Одноклассники», *Telegram,* «Дзен», *LiveJournal,* «МирТесен», *mt.gazeta.ru,* «Внутренняя политика России. Политический форум» и др, – всего 53 площадки.

Библиография

Багдасарян В. Э., Володенков С. В., Жмурин И. Е. Человек и технологический прогресс: антропологическая повестка мирового развития: моногр. Ярославль: Шукаева и семья, 2025.

Бродовская Е. В., Домбровская А. Ю., Пырма Р. В., Азаров А. А. Информационные потоки о мигрантах и для мигрантов в социальных медиа России // Информационное общество. 2020. № 6. С. 7–23.

Волков Ю. Г., Кривопусков В. В., Курбатов В. И. Цифровые мигранты и цифровая диаспора: новые проблемы и тренды международной миграции // Цифровая социология / DigitalSociology. 2021. Т. 4. № 4. С. 102–108.

Володенков С. В. Total Data как феномен формирования политической постреальности // Вестн. Омск. ун-та. Сер.: Исторические науки. 2017. № 3 (15). С. 409–415.

Глущенко Г. И. Развитие виртуальной миграции в контексте цифровизации // ДЕМИС. Демографические исследования. 2021. Т. 1. № 2. С. 57–64.

Гребенюк А. А., Субботин А. А. Исследование миграционных процессов в электронных социальных сетях // Цифровая социология / DigitalSociology. 2021. Т. 4. № 2. С. 23–31.

Искусных О. Интернет как фактор распространения эмиграционных установок среди молодёжи // Когнитивные науки в информационном обществе. 2021. Т. 1. № 2.

Aroles J., Bonneau C., Bhankaraully S. (2022) Conceptualising 'Meta-Work' in the Context of Continuous, Global Mobility: The Case of Digital Nomadism. *Work, Employment and Society* 37: 1261–1278.

Benoit K. (2020) "Text as Data: An Overview." In L. Curini, R. Franzese (eds.) *The SAGE Handbook of Research Methods in Political Science and International Relations*. London: SAGE Publications. Pp. 461–497.

Bronitsky G. (2024) Migration nowcasting using Google Trends: cross-country application. *Population and Economics* 8 (2):133–154.

Cook D. (2023) What is a digital nomad? Definition and taxonomy in the era of mainstream remote work. *World Leisure Journal* 6: 256–275.

Hochreiter S., Schmidhuber J. (1997) Long Short-Term Memory. *Neural Computation* 9 (8): 1735–1780.

Le Mens, G. and Gallego, A. (2025) Positioning Political Texts with Large Language Models by Asking and Averaging. *Political Analysis* 33 (3): 274–282. DOI 10.1017/pan.2024.29

Pu T., Huang Ch., Zhang H., Yang J., Huang M. (2024) Application of deep learning model incorporating domain knowledge in international migration forecasting. *Data Technologies and Applications* 58 (5): 787–806.

Welch N. G., Raftery A. E. (2022) Probabilistic forecasts of international bilateral migration flows. *Proceedings of the National Academy of Sciences of the United States of America* 119 (35): 1–8.

References

Aroles J., Bonneau C., Bhankaraully S. (2022) Conceptualising 'Meta-Work' in the Context of Continuous, Global Mobility: The Case of Digital Nomadism. *Work, Employment and Society* 37: 1261–1278.

Benoit K. (2020) "Text as Data: An Overview." In L. Curini, R. Franzese (eds.) *The SAGE Handbook of Research Methods in Political Science and International Relations*. London: SAGE Publications. Pp. 461–497.

Brodovskaya E. V., Dombrovskaya A. Yu., Pyrma R. V., Azarov A. A. (2020) Informatsionnye potoki o migrantakh i dlya migrantov v sotsial'nykh media Rossii [Information flows about and for migrants in Russian social media]. *Informatsionnoe obshchestvo* 6: 7–23. (In Russian)

Bronitsky G. (2024) Migration nowcasting using Google Trends: cross-country application. *Population and Economics* 8 (2):133–154.

Bagdasaryan V. E., Volodenkov S. V., Zhmurin I. E. (2025) *Chelovek i tekhnologicheskiy progress: antropologicheskaya povestka mirovogo razvitiya: monografiya* [Humankind and technological progress: an anthropological agenda for global development: monograph]. Yaroslavl: Shukaeva i sem'ya Publ. (In Russian)

Cook D. (2023) What is a digital nomad? Definition and taxonomy in the era of mainstream remote work. *World Leisure Journal* 6: 256–275.

Glushchenko G. I. (2021) Razvitie virtual'noy migratsii v kontekste tsifrovizatsii [The development of virtual migration in the context of digitalization]. *DEMIS. Demograficheskie issledovaniya* 1 (2): 57–64. (In Russian)

Grebenyuk A. A., Subbotin A. A. (2021) Issledovanie migratsionnykh protsessov v elektronnukh sotsial'nykh setyakh [Research on migration processes in online social networks]. *Tsifrovaya sotsiologiya / DigitalSociology* 4 (2): 23–31. (In Russian)

Hochreiter S., Schmidhuber J. (1997) Long Short-Term Memory. *Neural Computation* 9 (8): 1735–1780.

Iskusnykh O. (2021) Internet kak faktor rasprostraneniya emigratsionnykh ustanovok sredi molodezhi [The Internet as a factor in the spread of emigration attitudes among youth]. *Kognitivnye nauki v informatsionnom obshchestve* 1 (2). (In Russian)

Le Mens, G. and Gallego, A. (2025) Positioning Political Texts with Large Language Models by Asking and Averaging. *Political Analysis* 33 (3): 274–282. DOI 10.1017/pan.2024.29

Pu T., Huang Ch., Zhang H., Yang J., Huang M. (2024) Application of deep learning model incorporating domain knowledge in international migration forecasting. *Data Technologies and Applications* 58 (5): 787–806.

Volkov Yu. G., Krivopuskov V. V., Kurbatov V. I. (2021) Tsifrovye migranty i tsifrovaya diaspora: novye problemy i trendy mezhdunarodnoy migratsii [Digital migrants and the digital diaspora: new problems and trends in international migration]. *Tsifrovaya sotsiologiya / DigitalSociology* 4 (4): 102–108. (In Russian)

Volodenkov S. V. (2017) Total Data kak fenomen formirovaniya politicheskoy postreal'nosti [Total Data as a phenomenon of political post-reality formation]. *Vestnik Omskogo universiteta. Seriya "Istoricheskie nauki"* 3 (15): 409–415. (In Russian)

Welch N. G., Raftery A. E. (2022) Probabilistic forecasts of international bilateral migration flows. *Proceedings of the National Academy of Sciences of the United States of America* 119 (35): 1–8.

Поступила в редакцию
05.06.2025