DOI: 10.19181/4m.2024.33.2.2
EDN: HARLRA

ИСПОЛЬЗОВАНИЕ БОЛЬШИХ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ ДЛЯ АНАЛИЗА ВНУТРЕННЕЙ МИГРАЦИИ НАСЕЛЕНИЯ

Максимова Анастасия Сергеевна
МГУ имени М.В. Ломоносова, Москва, Россия
lubijizn@yandex.ru
ORCID: 0000-0003-3847-1791

Гребенюк Александр Александрович
МГУ имени М.В. Ломоносова, Москва, Россия
gaa-mma@mail.ru
ORCID: 0000-0001-9003-4551

Алешковский Иван Андреевич
МГУ имени М.В. Ломоносова, Москва, Россия
aleshkovski@yandex.ru
ORCID: 0000-0001-9276-3133

Для цитирования: *Максимова А. С., Гребенюк А. А., Алешковский И. А.* Использование больших данных социальных сетей для анализа внутренней миграции населения // Социология: методология, методы, математическое моделирование (Социология: 4М). 2024. № 59. С. 31-55. DOI: 10.19181/4m.2024.33.2.2. EDN: HARLRA

Статья посвящена разработке методики исследования миграционного движения населения на основе анализа больших данных социальных сетей через поиск закономерностей отражения процесса внутренней миграции в сообщениях пользователей социальных медиа. Проведенное исследование позволило оценить степень валидности и релевант-

ности такого рода цифровых следов индивидов как источника эмпирических данных о внутренней миграции. Для формирования базы исходных эмпирических данных, состоящей из сообщений о переезде, опубликованных пользователями социальных сетей, использовалось программное решение в виде платформы Brand Analytics. В результате апробации методики была установлена фрагментарность исследовательских, аналитических и прогностических возможностей использования сообщений в социальных сетях в качестве источника данных о внутрироссийской миграции населения. В качестве перспективы развития подобных исследований предложен нарративный анализ на основе сформированной выборки сообщений пользователей, имеющих опыт переселения внутри страны.

Ключевые слова: большие данные, социальные сети, социальные медиа, внутренняя миграция, источники данных о миграции, анализ сообщений, анализ миграции по данным социальных сетей.

Благодарности: исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Математические методы анализа сложных систем».

Введение

Исследование больших данных электронных социальных сетей – один из наиболее быстро развивающихся цифровых методов анализа сложных социальных систем. Возможность получать информацию об активности человека в виртуальном пространстве значительно расширила инструментарий социального аналитика. За короткое время сформировалась новая отрасль теоретической и практической деятельности – цифровая социология [1]. Концептуализировала этот термин и ввела в широкий научный оборот Д. Луптон в 2012 году, обозначив тем самым зарождение нового научного направления и появление новой методологии [2].

На сегодняшний день в мире опубликованы сотни научных работ, открыты десятки научных центров по данной проблема-

тике. При этом перечень цифровых методов и методик изучения социальных процессов постоянно расширяется ввиду перманентного развития информационных технологий [3]. Согласно данным агентства «We Are Social» и сервиса для SMM Hootsuite в 2022 году в мире количество пользователей социальных сетей насчитывало около 4,62 млрд. человек [4]. В России ежемесячное количество сообщений в сети «ВКонтакте» превышает 400 миллионов, а количество пользователей стремится к 25 миллионам. Согласно исследованию ВЦИОМ, в последние годы роль социальных сетей и социальных медиа достигла впечатляющих масштабов: 42% опрошенных респондентов отметили, что узнают новости экономики и общественно-политической жизни страны и своего региона из социальных сетей и блогов [5].

Ежедневно социальные сети формируют значительные по объемам данные, характеризующие различные социальные процессы. В этой связи сегодня особую актуальность приобретает задача использования этих данных с целью получения новых знаний об обществе, частным случаем которых является новый вектор развития исследований миграционных процессов. Отсутствие полных, достоверных и сопоставимых данных о перемещениях населения традиционно является значимым ограничением ее всестороннего изучения [6]. Так, сравнение официальной российской статистики об эмиграции с данными зарубежных стран об иммиграции российских граждан выявило их существенное расхождение [7]. Еще в большей степени ограниченность статистических данных характерна для внутригосударственных перемещений населения, отличающихся зачастую незадокументированной сменой места постоянного проживания, что определяет актуальность использования больших данных социальных сетей для анализа внутренней миграции.

Наличие вышеупомянутых проблем статистического наблюдения и появление возможности использования больших данных для анализа перемещений населения дало старт целому

ряду исследований различных видов миграции, проводимых как национальными научными центрами, так и международными организациями. Так, например, в 2017 году был опубликован доклад «Social Media and Forced Displacement: Big Data Analytics and Machine Learning: White Paper», посвященный использованию данных социальных сетей для анализа вынужденной миграции населения [8]. В 2022г. группой экспертов, сформированной Бюро Конференции европейских статистиков, был опубликован доклад «Использование новых источников данных для формирования статистики миграции», посвященный использованию новых источников данных для измерения международной миграции и трансграничной мобильности [9]. Изучению вынужденной миграции на основе данных социальных сетей посвящены работы таких зарубежных исследователей, как М. Александер [10], С. Виттеборн [11], Е. Гуалда [12], Н. Маркез [13], А. Рихи [14]. Немаловажное направление использования данных социальных сетей связано с изучением интеграции и ассимиляции мигрантов. А. Дубуа и соавторы на основании анализа «лайков» в социальных сетях разработали методику анализа ассимиляции говорящих по-арабски мигрантов, проживающих в Германии [15]. И. Стюард проанализировал интеграцию мигрантов из Мексики в США на основе анализа прослушиваемых в социальных сетях музыкальных композиций [16].

В российской миграциологии накоплен богатый опыт изучения внутренней миграции населения. Вместе с тем в отечественных исследованиях уделяется недостаточное внимание цифровым методам изучения этого социального процесса. Среди немногих исключений – результаты реализации проекта «Виртуальное население России» Н.Ю. Замятиной и А.Д. Яшунского – базы данных по структуре крупнейшей российской социальной сети «ВКонтакте». Предложенную авторами методику можно использовать для анализа внутрироссийских перемещений молодежи [17]. В работе К.А. Чернышева на основе анализа цифровых

данных социальной сети «ВКонтакте» осуществлены расчёты коэффициентов интенсивности миграционных связей населения Крыма [18].

Целью настоящего исследования являлось выявление закономерностей отражения процесса мобильности населения из регионов России по направлению в Москву и Московскую область в сообщениях пользователей социальных сетей. Кроме того, ввиду отсутствия аналогичных исследований необходимо было сформировать методологическую основу исследований миграционного движения на основе текстовых данных социальных сетей, а также оценить степень их валидности и релевантности как источника информации для обогащения эмпирической информации о внутренней миграции и конкретизировать исследовательские задачи, которые помогает решить данный источник.

Методология и методика исследования

Эмпирический объект исследования составила совокупность сообщений в социальных сетях о переезде их авторов из регионов России для постоянного проживания и/или трудоустройства в г. Москву и Московскую область. Анализу не подлежали сообщения, в которых речь шла о переезде не самого автора, а третьих лиц. Таким образом, эмпирический объект сводился к собственному опыту переезда, о котором сообщал тот или иной пользователь сети посредством публикации сообщения. В анализируемый массив были включены посты и комментарии, исключены репосты без комментирования, поскольку это создает двойной счет, и противоречит основной парадигме сбора данных, уже упомянутой выше: «сам автор сообщает о своем переезде».

Методика работы с эмпирическими данными состояла из двух этапов: на первом производился сбор данных и стандартные процедуры предобработки, позволяющие улучшить каче-

ство обрабатываемой в последующем информации, на втором этапе проводился анализ полученной информации.

Содержательные задачи первого этапа сводились к:

– разработке лингвистического запроса (набора ключевых слов, по которым идентифицируются сообщения о миграции, и минус-слов, исключающих сообщения из массива) для формирования массива данных;

– автоматической очистке данных (удаление дубликатов, спама, сообщений, содержащих только изображения, репосты чужих сообщений без добавления собственного комментария, исторических очерков, биографий известных людей, анекдотов и иных сообщений, которые не относятся к изучаемой теме);

– ручная оценка релевантности собранной на первом этапе информации поставленным задачам на основе плотности распределения в 10%-ной подвыборке.

Таким образом, был сформирован массив сообщений с присущими им атрибутами: дата публикации и информация страницы автора сообщения, поступившего в обработку.

Алгоритм работы с полученной из социальной сети эмпирической социологической частично структурированной информацией на втором этапе реализации исследования состоял из следующих шагов:

– разработка концепции качественного анализа неструктурированной части массива полученных данных с точки зрения цели и проблемного поля исследования;

– количественный анализ распределения атрибутов сообщений и их моделирование;

– содержательный анализ текстов сообщений.

По итогам реализации первого этапа работы с данными была получена совокупность сообщений из социальных сетей, в которых индивиды сообщают о собственном переезде в Москву или Московскую область, вне зависимости от того, когда произошел этот переезд.

Необходимо отметить, что полнота собранных данных для созданного запроса не может быть оценена, поскольку за пределами выборки остаются сообщения, не вошедшие в исходный лингвистический запрос. Полноту сбора ограничивает тот факт, что переезд может быть отражен в сообщении достаточно неявно и без обозначения ключевых слов, из которых состоял лингвистический запрос. При оценке по 10%-ной случайной подвыборке было установлено, что несмотря на автоматическую обработку, на 1800 сообщений присутствуют 468 сообщений, или 26%, не соответствующих парадигме отбора (сообщение автором о своем переезде), однако, релевантных цели исследования. В основном это фрагменты художественных произведений и биографические записки об известных людях, письма в редакцию, интервью и различные варианты рерайта этих же интервью. Именно интервью и цитаты, содержащие ключевые фразы запроса о собственном переезде, составляют основную сложность при их фильтрации для соответствия эмпирическому объекту исследования. Поэтому была произведена ручная фильтрация сообщений для текстового анализа, однако количественный анализ был произведен по полному массиву данных, полученных в результате выгрузки. Таким образом, количественный анализ представляет собой характеристику присутствия темы миграции в информационном поле, а качественный – позволяет получить выводы об отражении авторами опыта собственного переезда в сообщениях.

Для сбора данных применялось готовое программное решение в виде платформы Brand Analytics [19], позволяющее получить архивные данные с ретроспективой не более года, поэтому в обработку поступили данные с 04.08.2022 г. по 04.08.2023 г. Сбор данных осуществлялся с личных страниц пользователей и страниц сообществ в тех социальных сетях, где они существуют. Ввиду ограничений приватности, установленных самими пользователями, сбор данных осуществлялся с тех страниц,

на которых подобных ограничений нет. Были собраны сообщения и комментарии, удовлетворяющие лингвистическому запросу.

Обработка и анализ данных производились на языке Python в среде Google Colaboratory. При работе с данными использовались методы анализа временного ряда, кластерный анализ, методы автоматической обработки естественного языка (NLP).

Результаты исследования

За анализируемый период (04.08.2022 г. – 04.08.2023 г.) было собрано более 17 тыс. сообщений. Однако, полученное число нельзя сопоставлять с численностью переселившегося за это же время населения, поскольку часть сообщений относится к опыту переезда, произошедшему задолго до времени публикации сообщения. При этом в некоторых сообщениях есть указание на время, когда был совершен переезд (например, «уже год как я живу в Москве»), однако унификация подобного указания, то есть составление описания всех возможных указаний времени и классификация сообщений по ним для последующего восстановления времени переселения, не являлась задачей настоящего исследования.

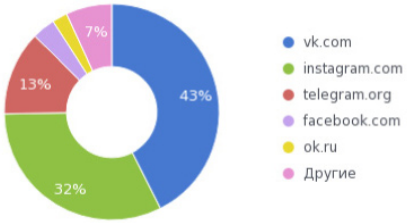

Рис.1. Распределение сообщений по наиболее используемым платформам публикации, в % от общей доли анализируемых сообщений [1]

[1] Социальные сети Instagram и Facebook запрещены на территории России.

Среди всех социальных сетей наибольшее количество публикаций было сделано в vk.com. Зачастую, один и тот же автор размещает идентичные сообщения сразу на нескольких платформах, что создает проблему двойного счета, однако удаление дубликатов позволяет частично решить эту проблему.

Для текстового анализа были выбраны сообщения из источников, в которых преобладают текстовые сообщения на платформах vk.com, ok.ru, twitter.com, facebook.com. Текстовые данные из Youtube были получены в виде тайм-кода, либо транскрипции, что делает их громоздкими и зашумленными. Кроме того, видео источник имеет иной спектр воздействия как на аудиальные центры восприятия информации, так и на визуальные, поэтому получаемая посредством него информация, вполне вероятно интерпретируется пользователями данного ресурса иначе, чем, например, текстовые сообщения из социальной сети ВКонтакте. Также следует отметить, что размещение видео на Youtube сложнее и не каждый пользователь видеохостинга публикует видео: большинство только смотрят. Поэтому аккаунты из Youtube, попавшие в выборку, имеют свои характерные черты, отличающие их от аккаунтов авторов в социальной сети ВКонтакте. С этой точки зрения социальные сети с преобладанием текстового контента носят иной характер доступности публикации собственной информации и получения размещаемой информации пользователями.

Общий количественный анализ в рамках исследования проводился по всем источникам, что позволило охарактеризовать общий фон всего информационного поля социальных сетей и онлайн СМИ с точки зрения присутствия в нем темы внутристрановой миграции по направлению в столичный регион.

С учетом общего количества авторов публикаций (более 15 тыс.), на каждого автора приходилось 1,18 сообщения. Такое значение показателя достигается за счет страниц сообществ (существующих в некоторых социальных сетях), на которых сооб-

щения о миграции могут быть опубликованы многократно. Среди них наиболее выделяются три сообщества, в которых были опубликованы от 18 до 33 сообщений, в остальных 20 сообществах с публикациями о переезде было не более 9 сообщений. Существуют также тематические сообщества, в которых их владельцы рассказывают о переезде в Москву, либо предлагают услуги трудоустройства. Однако, подобная форма контента популярна и характерна больше для релокации в другую страну, поскольку имеются межстрановые культурные и социально-экономические различия, которыми интересуется аудитория. В случае с переездом в границах одной страны, подобных различий нет, и подобный контент не настолько востребован.

Основная тематика сообщений – автобиографические записи, в повествовательном стиле рассказывающие о жизни автора, в которых переезд в Москву становится либо центральной темой и переломным моментом в жизни, либо о нем сообщается вскользь в контексте другого более эмоционально окрашенного события («пост-знакомство.. Я переехала из маленького села, в Московскую область…», либо «это был шок»). Большая доля сообщений связана с этапом адаптации: поиском квартиры и друзей в новом месте проживания.

Общая аудитория страниц, на которых были опубликованы сообщения о переселении, составила 965,2 тыс. аккаунтов. Этот показатель предполагает потенциальное количество пользователей, прочитавших данные сообщения. Количество реальной аудитории оценить достаточно сложно, поскольку часть страниц, входящих в потенциальную аудиторию, не используется, а часть принадлежит не физическому лицу, а ведется редакторскими группами, либо это может быть страница, созданная с целью продвижения коммерческих услуг, или иные формы существования страниц, предполагающие, что их ведет пользователь, не являющийся реальным физическим лицом, способным формировать социальную реальность и общественное мнение. В данном

случае разработка подобных критериев, является отдельной задачей. Проанализировать количество просмотров не позволяет поисковая система, поскольку не агрегирует такую информацию.

Средняя вовлеченность (суммарное количество всех реакций в расчете на одно сообщение) составила 182 за весь период, при этом наиболее комментируемыми сообщениями были сообщения аккаунтов с большими количественными атрибутами – размером аудитории. Однако общий внешний инфоповод в них выделить не удалось.

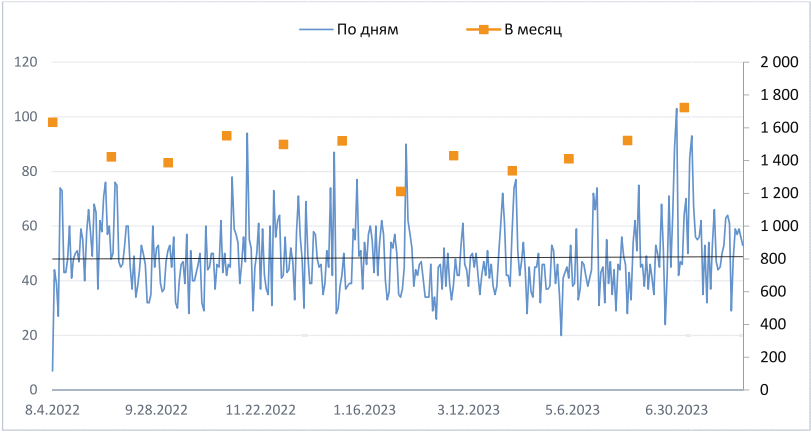

Рис.2. Динамика сообщений о переезде по датам (основная ось, линия) и кумулятивно за каждый месяц (вспомогательная ось, точки), шт.

В течение исследуемого периода каждый месяц было опубликовано около 1,5 тыс. сообщений кумулятивным итогом. Визуально процесс выглядит как стационарный (рис.1) ввиду постоянства среднего значения, постоянства дисперсии и отсутствия тренда (угол наклона нулевой, аппроксимация основными типами функций дает значения R^2 близкое к 0).

Сезонность публикаций на основе автокорреляций периода до 30 не выявлена (рис.2). Первые два периода (с лагом 1 и лагом 2) были проверены, исходя из теоретической гипотезы о том, что начало обсуждения темы – пост, влечет за собой комментарии, в которых могут наблюдаться слова и словосочетания, присущие исходному посту. Однако проведенное исследование выявило, что тема внутренней миграции не настолько обсуждаема, чтобы пост провоцировал еще ряд постов на эту же тему. Тест Дики-Фулера позволяет отвергнуть нулевую гипотезу о нестационарности временного ряда (ADF Statistic: - 4.226, p-value: 0.00).

Наиболее заметные скачки количества сообщений, наблюдаются в даты 15.11.2022 г., 04.01.2023 г., 07.02.2023 г., 04.07.2023 г., 08.07.2023 г. (рис.2).

В ходе исследования было установлено, что подобный рост не связан с внешними факторами, поскольку сообщения не имеют общего инфоповода, то есть рост является случайностью за исключением одной экстремальной точки, которая наблюдалась 31 декабря. Подобное пиковое значение связано с рефлексией пользователей о прошедших в их жизни событиях за уходящий год («итак, я выполнила 71 цель из поставленных», «подходит к концу 2022 год, и я хочу..», «новый год. За этот год я переехал из Рязани в Москву»). За весь период наблюдался один наиболее явный пик вовлеченности, что было связано с публикацией сообщения о переезде в Москву блогером, имеющим аудиторию более 600 тыс. человек.

Таким образом, исследование показало, что совокупность сообщений в социальных сетях о переезде является набором случайных величин, не детерминированных явными внешними факторами, и имеет постоянство характеристик временного ряда. Иными словами, ежедневно пользователи размещают около 50 постов в социальной сети о собственном переезде вне зависимости от текущих внешних факторов и флуктуаций информационного поля. Детерминированный рост активности, в отношении

которого может быть выдвинута гипотеза о влиянии соответствующего фактора, наблюдается только в последний день уходящего года. Однако, период наблюдения, равный одному году, не может свидетельствовать о повторяемости данного процесса и в другие годы, для этого необходимо наблюдение за более длительный период.

Личная информация, указанная пользователями о себе, может быть использована для описания социально-демографических характеристик совокупности внутренних мигрантов, однако, в разных социальных сетях пользователи имеют возможность заполнить различную информацию о себе. Таким образом, атрибуты страниц в разных социальных сетях не всегда совпадают. Среди личных профилей, опубликовавших сообщения о переезде, только на 53% страниц указан пол, на 14% – возраст, что не позволяет охарактеризовать социально-демографический портрет пользователей, публиковавших сообщения о переезде. На основе имеющихся данных восстановить пропущенные значения возможно только в случае распределения неизвестных величин аналогичного известным величинам без смещения, либо с заранее известным смещением. Например, если будет точно известно о том, что молодежь не пишет о своем возрасте, а с увеличением значения возраста, пользователи склонны его указывать. Причем смещения полученной оценке социально-демографического портрета добавят намеренные недостоверные сведения о половозрастной принадлежности, доля которых не может быть установлена.

Анализ географического положения аккаунтов в данном случае также не информативен, поскольку анализируемое направление переезда в большинстве случаев определяет текущее географическое положение: у большинства аккаунтов география определяется как г. Москва. Место выхода переселенца также возможно определить только в случае его указания в тексте сообщения.

Из описанных ограничений следует, что текстовая информация в социальных сетях, анализ которой доступен исследователям в настоящее время, не позволяет получать обоснованные количественные оценки внутренней миграции. В силу высокой степени нечеткости, неполноты и высокой вариативности, такая информация не может использоваться ни для построения показателей, альтернативных по отношению к официальной статистике, ни для верификации последней. Тем не менее, большие масштабы доступного текста сообщений в соцсетях могут делать их ценным источником информации о переживаниях и смыслах, связанных с таким социальным явлением как миграция. Поэтому далее предложены перспективы развития аналитических исследований в этом направлении.

Одним из возможных аспектов анализа является определение тональности сообщений, поскольку это поможет понять отношение к переезду и сложности в процессе. Система Brand Analytics принимает на вход указание объекта, по отношению к которому будет определяться тональность. Автоматическое определение эмоциональной окраски по отношению непосредственно к переезду не дает существенного результата, поскольку 98% сообщений определяются системой как нейтральные. Это исключает идентификацию двойной тональности, когда в одной части предложения эмоциональная окраска положительная, а в другой отрицательная. Однако, проблема заключается в том, что как таковой переезд, о котором идет речь в сообщениях, пользователями, как правило, не сопровождается прилагательными, позволяющими присвоить ему тональность. Как существительное слово «переезд» чаще всего отсутствует, таким образом отсутствует объект определения тональности. Альтернативным является подход, определяющий тональность по отношению к месту предыдущего проживания, либо к Москве.

Также существует возможность автоматически определить тональность на основе машинного обучения. Для этого необхо-

дима тренировочная выборка, представляющая собой предварительно размеченную часть анализируемого массива на позитивные и негативные сообщения, что является трудозатратным процессом, либо готовые размеченные наборы данных.

Рис. 3. Облако слов по частоте встречаемости в сообщениях о миграции (величина шрифта пропорциональна частоте встречаемости)

В рамках исследования была также проанализирована содержательная часть сообщений. Результаты анализа встречаемости представлены в виде облака тегов, центральное место среди которых занимает «Москва», наиболее значимые позиции занимают слова: «жизнь», «время», «годы», «день», далее следуют слова, обозначающие составляющую повседневной жизни: «работа», «люди», «город», «друг», «дом», «место». Таким образом, основное содержание сообщений связано с целью и обстоятельствами переезда, материальными условиями адаптации после переезда, его ролью в жизни человека (см. рис. 3). Однако такой анализ, оценивая частотность слов, не позволяет выделять основные смыслы, поскольку не учитывает совместную встречаемость выделенных наиболее часто встречающихся слов. Поэтому была произведена кластеризация и тематическое обобщение текстов

с помощью библиотек, позволяющих работать с текстовыми данными на языке Python.

Задача тематического моделирования была решена на данных, предварительно подготовленных с помощью разделения текста на слова («токенизации»), извлечения их базовой формы («лемматизации») и удаления стоп-слов [20]. Далее данные были векторизованы, то есть переведены в числовой вид. Общий алгоритм работы с данными представлен на рисунке 4.

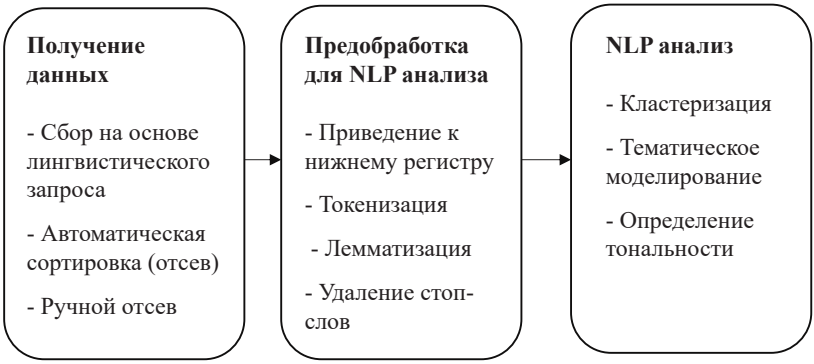

Рис. 4. Процесс работы с текстовыми данными, использованными в исследовании

В рамках настоящей статьи представлены результаты всех шагов за исключением определения тональности.

С целью разбиения сообщений по темам был применен метод обучения без учителя ввиду того, что разметка данных на примере с сообщениями о миграции является трудоемкой. Для первичной апробации из сформированного случайным образом ряда сообщений были вручную отобраны первые 1000 сообщений, попадающих под основные условия отбора (сам пользователь сообщает о своем переезде). Далее была проведена кластеризация методом k - средних с подбором на основе «метода

локтя» оптимального количества кластеров, и для сообщений, попавших в один кластер, была выделена общая тема с помощью алгоритма сжатия текста. Для обработки данных использовалась библиотека PyTorch.

Оптимальным количеством было признано 4 кластера сообщений, частоты которых на первой и второй итерации запуска модели были распределены практически идентично. В силу схожести основных текстовых обобщений сообщений в кластерах, были выделены темы, наиболее обсуждаемые в анализируемых сообщениях, однако, не для всех кластеров удалось их скомпилировать. Было установлено, что основные тематические классы сообщений о переезде, посвящены 1) самопрезентации (так называемые «посты-знакомства», например: «Познакомимся с руководителем компании»), 2) рефлексии и осмыслению опыта собственной жизни («Мне нравится переезд в Москву из Орла», «Спасибо за каждый прожитый день», «В очередной раз я порассуждаю о моей наболевшей теме»), в которых описывается то, какую роль сыграл переезд, и 3) описанию непосредственно тонкостей процесса переезда («Что нужно было решить до переезда в Москву?»). Поскольку кластеризация и тематическое моделирование были осуществлены на подвыборке, есть вероятность, что некоторый миноритарный, редко встречающийся кластер, не вошел в подвыборку и не был обработан.

Заключение

В рамках настоящего исследования был разработан поисковый запрос в виде словаря для поиска в социальных сетях сообщений о переезде из регионов России в Москву и Московскую область. Словарь может быть расширен для идентификации сообщений о других направлениях переселения, а также использован при дальнейшем развитии исследований, посвященных исследованию миграции по данным сообщений в социальных медиа.

Проведенное исследование позволило сделать вывод, что, несмотря на большой объем доступных текстовых данных о внутрироссийской миграции в Москву и Московскую область, сообщения в социальных сетях не являются тем источником, который может верифицировать имеющиеся или дать альтернативные количественные оценки внутренней миграции. Поскольку отсутствует возможность сопоставления обнаруженных в социальной сети сообщений о миграционных перемещениях, являющихся распределенными относительно акта миграции, который описывается в них, и данных официального статистического учета внутрироссийских перемещений, невозможно также производить оценку колебаний явления, в том числе сезонных. Таким образом, в текстах социальных медиа наблюдается только отражение части реального миграционного опыта населения, распределенное неупорядоченным образом относительно реальной хронологии событий.

Вторая существенная проблема – достоверность информации. При работе с текстовыми массивами сообщений была обнаружена большая доля художественных произведений, упоминаний о переезде в контексте рекламы собственных услуг и т.д., а также сообщений о переезде третьих лиц, которые несут еще больше субъективизма. Тем самым для анализа мотивов переезда и проблем при последующей приживаемости в новом месте, необходим подбор сообщений, в которых переселение является центральной темой, и их подробное исследование.

Было выявлено, что сообщения о переезде в Москву не провоцируют дискуссий в комментариях, а представляют собой автобиографические записки, в которых авторы делятся собственным опытом, что не вызывает бурной реакции аудитории в виде текстового обсуждения обстоятельств переезда. В зависимости от основного смыслового содержания сообщений, среди них можно выделить три группы: сообщения, в которых пользователь рассказывает о своей жизни («сообщения-знакомства»), сообщения,

в которых пользователь рефлексирует о своем опыте в жизни и сообщения, в которых описывается процесс переезда.

Таким образом, необходимо подчеркнуть фрагментарность исследовательских, аналитических и прогностических возможностей использования текстовых сообщений в социальных сетях в качестве источника информации о внутрироссийской миграции.

В качестве перспектив развития исследования рассматривается нарративный анализ на основе сформированной выборки сообщений, принадлежащих пользователям, имеющим опыт переселения, а также использование больших языковых моделей (например, ChatGPT, GigaChat) для автоматизации задач NLP. Последнее должно позволить снизить трудоемкость задач подготовки размеченной выборки для определения тональности, а также улучшить фильтрацию сообщений, не отвечающих требуемой теме анализа. Все это позволит осуществлять социальные исследования того опыта миграции, который отражен в большом массиве текстовых сообщений на соответствующую тему, имеющихся в социальных сетях.

ЛИТЕРАТУРА

1. Шульц В.Л., Гребенюк А.А., Ашманов И.С. Теоретико-методологические проблемы цифровой социологии // Вестник Московского университета. Серия 18. Социология и политология. 2022, т. 28, № 1. С. 126-144. DOI: 10.24290/1029-3736-2022-28-1-126-144. EDN: SWVTCX.

2. Lupton D. Digital Sociology: An Introduction. Sydney: University of Sydney, 2012. 17 p. DOI: 10.2139/ssrn.2273418.

3. Орлова И.Б., Фомин Е.В. Цифровая социология: возможности, риски, перспективы // Национальная безопасность/Nota Bene. 2020, № 3. С. 48-63. DOI: 10.7256/2454-0668.2020.3.33274. EDN: MJWSXZ.

4. Digital 2022: Another year of bumper growth // We Are Social: [сайт]. 26.01.2022. URL: https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/ (дата обращения: 27.12.2024).

5. Исследование ВЦИОМ «Медиапотребление и активность в интернете» // ВЦИОМ: [сайт]. 23.09.2021. URL: https://wciom.ru/analytical-reviews/analiticheskii-obzor/mediapotreblenie-i-aktivnost-v-internete (дата обращения: 27.12.2024).

6. Гребенюк А.А., Субботин А.А. Исследование миграционных процессов в электронных социальных сетях // Цифровая социология. 2021, т. 4, № 2. С. 23-31. DOI: 10.26425/2658-347X-2021-4-2-23-31. EDN: FVKNMY.

7. Aleshkovski I., Gasparishvili A., Grebenyuk A. The Changing Landscape of Russia's Emigration from 1990 to 2020: Trends and Determinants // Journal of Globalization Studies. 2023, vol. 14, № 1. P. 42–65. DOI: 10.30884/jogs/2023.01.04

8. Social Media and Forced Displacement: Big Data Analytics and Machine Learning: White Paper. // UN Global Pulse and UNHCR Innovation Service. 09.2017. URL: https://www.unhcr.org/innovation/wp-content/uploads/2017/09/FINAL-White-Paper.pdf (дата обращения: 27.12.2024).

9. Использование новых источников данных для формирования статистики миграции (2022). Заседание Группы экспертов по статистике миграции. 26-28 октября 2022 года, г. Женева. Рабочий документ 15 // UNECE. URL: https://unece.org/sites/default/files/2022-1/WP15_TaskForce_NewDataSourcesMigration_RUS.pdf (дата обращения: 27.12.2024).

10. Alexander M., Polimis K., Zagheni E. The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data // Population and Development Review. 2019, vol. 45, № 3. P. 617–630. DOI: 10.1111/padr.12289.

11. Witteborn S. The digital gift and aspirational mobility // International Journal of Cultural Studies. 2019, vol. 22, № 6. P. 754–769. DOI: 10.1177/1367877919831020.

12. Gualda E., Rebollo C. The refugee crisis on Twitter: A diversity of discourses at a European crossroads // Journal of Spatial and Organizational Dynamics. 2016, vol. 4, № 3. P. 199–212.

13. Marquez N., Garimella K., Toomet O., Weber I.G., Zagheni E. Segregation and sentiment: Estimating refugee segregation and its effects using digital trace data // Guide to Mobile Data Analytics in Refugee Scenarios: The Data for Refugees Challenge Study / Ed. by A. Salah, A. Pentland, B. Lepri, E. Letouzé. Cham: Springer, 2019. P. 265–282. DOI: 10.1007/978-3-030-12554-7_14.

14. Righi A. Assessing migration through social media: a review // Mathematical Population Studies, 2019, vol. 26, № 2. P. 80–91. DOI: 10.1080/08898480.2019.1565271.

15. Dubois A., Zagheni E., Garimella K., Weber I. Studying migrant assimilation through Facebook interests // Social Informatics. SocInfo 2018. Lecture Notes in Computer Science, V. 11186 / Ed. by S. Staab, O. Koltsova, D. Ignatov. Cham: Springer, 2018. P. 51–60. DOI: 10.1007/978-3-030-01159-8_5.

16. Stewart I., Flores R., Riffe T., Weber I., Zagheni E. Rock, Rap, or Reggaeton?: Assessing Mexican immigrants' cultural assimilation using Facebook data // Proceedings of the World Wide Web Conference (WWW '19). San Francisco, CA, USA, May 13–17, 2019 / Ed. by L. Liu, R. White. New York, USA: Association for Computing Machinery, 2019. P. 3258–3264. DOI: 10.1145/3308558.3313409.

17. Замятина Н. Ю., Яшунский А. Д. Виртуальная география виртуального населения // Мониторинг общественного мнения: экономические и социальные перемены. 2018, № 1. С. 117—137. DOI: 10.14515/monitoring.2018.1.07. EDN: YQUCNL.

18. Чернышев К.А., *Чернышева Н.В., Петров. Е. Ю.* Межрегиональные связи населения Крыма: исследование на основе цифровых и статистических данных о местах рождения мигрантов // Геополитика и экогеодинамика регионов. 2022, т. 8, № 3. С. 253-264. EDN: EXIRYT.

19. Brand Analytics – российская система сбора и анализа данных социальных медиа: [сайт]. URL: https://br-analytics.ru/ (дата обращения: 27.12.2024).

20. Bird S., Klein E., Loper E. Natural Language Processing with Python. Sevastopol: O'Reilly Media, 2009. 502 p. ISBN: 978-0-596-51649-9.

Сведения об авторах

Максимова Анастасия Сергеевна
Кандидат экономических наук, доцент Высшей школы современных социальных наук МГУ имени М.В. Ломоносова
SPIN-код: 7343-4140
AuthorID: 729354

Гребенюк Александр Александрович
Доктор экономических наук, заместитель директора ВШССН МГУ имени М.В. Ломоносова
SPIN-код: 4007-9651
Scopus ID: 56297122600

Алешковский Иван Андреевич
Кандидат экономических наук, доцент Факультета глобальных процессов МГУ имени М.В. Ломоносова
Scopus ID: 57190586626

DOI: 10.19181/4m.2024.33.2.2

LEVERAGING SOCIAL MEDIA BIG DATA TO ANALYZE INTERNAL MIGRATION

Maksimova Anastasia S.
Lomonosov Moscow State University, Moscow, Russian Federation
lubijizn@yandex.ru
ORCID: 0000-0003-3847-1791

Grebenyuk Alexander A.
Lomonosov Moscow State University, Moscow, Russian Federation
gaa-mma@mail.ru
ORCID: 0000-0001-9003-4551

Aleshkovski Ivan A.
Lomonosov Moscow State University, Moscow, Russian Federation
aleshkovski@yandex.ru
ORCID: 0000-0001-9276-3133

For citation: Maksimova A. S., Grebenyuk A. A., Aleshkovski I. A. Leveraging social media big data to analyze internal migration. *Sotsiologiya: 4M (Sociology: methodology, methods, mathematical modeling),* 2024, no. 59, p. 31-55. DOI: 10.19181/4m.2024.33.2.2. EDN: HARLRA.

Abstract. The article is devoted to the development of a methodology for the study of population migration based on the analysis of big data in social networks through the search for patterns showing the internal migration process in the messages of social media users. The research in question allowed us to evaluate the degree of validity and relevance of digital traces of individuals as a source of empirical data on internal migration.

A software solution in the form of the Brand Analytics platform was used to generate the initial empirical data base onsisting of relocation messages published by social media users. The approbation of the methodology showed the fragmentation of research, analytical and predictive possibilities of using social networks as a source of data on intra-Russian population migration.

Narrative analysis based on the formed sample of messages of users who have experience of resettlement within the country was proposed for further development of similar types of research.

Keywords: big data, social networks, social media, internal migration, migration data sources, analysis of social media posts, migration analysis based on social media data.

Acknowledgments: the study was carried out with the support of the Interdisciplinary Scientific and Educational School of Moscow University "Mathematical Methods for the Analysis of Complex Systems."

References

1. Shul'c V.L., Grebenyuk A.A., Ashmanov I.S. Theoretical and methodological problems of digital sociology (in Russian), *Bulletin of the Moscow University. Series 18. Sociology and political science*, 2022, vol. 28, no 1, p. 126-144. DOI: 10.24290/1029-3736-2022-28-1-126-144. EDN: SWVTCX.
2. Lupton D. Digital Sociology: An Introduction. Sydney: University of Sydney, 2012, 17 p. DOI: 10.2139/ssrn.2273418.
3. Orlova I.B., Fomin E.V. Digital Sociology: opportunities, risks, prospects (in Russian), *National security. Nota Bene,* 2020, no 3, p. 48-63. DOI: 10.7256/2454-0668.2020.3.33274.
4. Digital 2022: Another year of bumper growth, *We Are Social:* [site]. 26.01.2022. URL: https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/ (date of access: 27 December 2024).
5. VTSIOM research "Media Consumption and Internet activity" (in Russian), *VTSIOM:* [site]. 23.09.2021. URL: https://wciom.ru/analytical-reviews/analiticheskii-obzor/mediapotreblenie-i-aktivnost-v-internete (date of access: 27 December 2024).
6. Grebenyuk A.A., Subbotin A.A. Research of migration processes in electronic social networks (in Russian), *Digital Sociology,* 2021, vol. 4, no 2, p. 23-31. DOI: 10.26425/2658-347X-2021-4-2-23-31. EDN: FVKNMY.
7. Aleshkovski I., Gasparishvili A., Grebenyuk A. The Changing Landscape of Russia's Emigration from 1990 to 2020: Trends and Determinants (in Russian), *Journal of Globalization Studies*, 2023, vol. 14, no. 1, p.42–65. DOI: 10.30884/jogs/2023.01.04.

8. Social Media and Forced Displacement: Big Data Analytics and Machine Learning: White Paper, *UN Global Pulse and UNHCR Innovation Service.* 09.2017. URL: https://www.unhcr.org/innovation/wp-content/uploads/2017/09/FINAL-White-Paper.pdf (date of access: 27 December 2024).
9. Using new data sources to generate migration statistics (2022) (in Russian), Meeting of the Expert Group on Migration Statistics, October 26-28, 2022, Geneva. Working paper 15, *UNECE.* URL: https://www.unhcr.org/innovation/wp-content/uploads/2017/09/FINAL-White-Paper.pdf (date of access: 27 December 2024).
10. Alexander M., Polimis K., Zagheni E. The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data, *Population and Development Review*, 2019, vol. 45, no. 3, p. 617–630. DOI: 10.1111/padr.12289.
11. Witteborn S. The digital gift and aspirational mobility, *International Journal of Cultural Studies*, 2019, vol. 22, no. 6, p. 754–769. DOI: 10.1177/1367877919831020.
12. Gualda E., Rebollo C. The refugee crisis on Twitter: A diversity of discourses at a European crossroads, *Journal of Spatial and Organizational Dynamics*, 2016, vol. 4, no. 3, p. 199–212.
13. Marquez N., Garimella K., Toomet O., Weber I.G., Zagheni E. Segregation and sentiment: Estimating refugee segregation and its effects using digital trace data, *Guide to Mobile Data Analytics in Refugee Scenarios: The Data for Refugees Challenge Study,* Springer, 2019, p. 265–282. DOI: 10.1007/978-3-030-12554-7_14.
14. Righi A. Assessing migration through social media: a review, *Mathematical Population Studies*, 2019, vol. 26, no. 2, p. 80–91. DOI: 10.1080/08898480.2019.1565271
15. Dubois A., Zagheni E., Garimella K., Weber I. Studying migrant assimilation through Facebook interests, *Social Informatics. SocInfo 2018. Lecture Notes in Computer Science*, V. 11186. Cham: Springer, 2018, p. 51–60. DOI: 10.1007/978-3-030-01159-8_5.
16. Stewart I., Flores R., Riffe T., Weber I., Zagheni E. Rock, Rap, or Reggaeton?: Assessing Mexican immigrants' cultural assimilation using Facebook data, *Proceedings of the World Wide Web Conference (WWW '19)*. San Francisco, CA, USA, May 13–17, 2019, ed. by L. Liu,

R. White. New York, USA: Association for Computing Machinery, 2019, p. 3258–3264. DOI: 10.1145/3308558.3313409.
17. Zamyatina N.YU., Yashunskij A.D. Virtual geography of the virtual population (in Russian), *Public opinion monitoring: economic and social changes,* 2018, no 1, p. 117—137. DOI: 10.14515/monitoring.2018.1.07. EDN: YQUCNL.
18. Chernyshev K.A., Chernysheva N.V., Petrov E.Yu. Interregional relations of the Crimean population: a study based on digital and statistical data on the places of birth of migrants (in Russian), *Geopolitics and ecogeodynamics of regions,* 2022, vol. 8, no 3, p. 253-264. EDN: EXIRYT.
19. *Brand Analytics – the Russian system for collecting and analyzing social media data:* [site]. URL: https://br-analytics.ru/ (date of access: 27 December 2024).
20. Bird S., Klein E., Loper E. Natural Language Processing with Python. Sevastopol: O'Reilly Media, 2009. 502 p. ISBN: 978-0-596-51649-9.

Information about the authors

Anastasia S. Maksimova
Ph.D. in Economics, Associate Professor, Department of Sociology of Knowledge, Higher School of Contemporary Social Sciences of the Lomonosov Moscow State University
SPIN-code: 7343-4140
AuthorID: 729354

Alexander A. Grebenyuk
Doctor of Economics, Deputy Director of the Higher School of Contemporary Social Sciences of the Lomonosov Moscow State University
SPIN-code: 4007-9651
Scopus ID: 56297122600

Ivan A. Aleshkovski
Ph.D. in Economics, Associate Professor, Faculty of Global Studies of the Lomonosov Moscow State University
Scopus ID: 57190586626