

Определение аргументации по вопросам репродуктивного поведения в мнениях пользователей социальных сетей методами обработки естественного языка

Ирина Е. Калабихина¹, Екатерина А. Зубова², Наталья В. Лукашевич³,
Антон В. Колотуша¹, Зарина Г. Казбекова¹, Евгений П. Банин⁴,
Герман А. Клименко¹

1 Экономический факультет МГУ имени М.В. Ломоносова, Москва, 119991, Россия

2 Йельский университет, Нью-Хейвен, Коннектикут, 06511, США

3 НИВЦ МГУ имени М.В. Ломоносова, Москва, 119899, Россия

4 НИЦ Курчатовский институт, Москва, 123182, Россия

Получено 2 November 2022 ♦ Принято в печать 10 May 2023 ♦ Опубликовано 30 June 2023

Цитирование: I Kalabikhina, E Zubova, N Loukachevitch, A Kolotusha, Z Kazbekova, E Banin, G Klimenko (2023) Identifying Reproductive Behavior Arguments in Social Media Content Users' Opinions through Natural Language Processing Techniques. Population and Economics 7(1):40–59. <https://doi.org/10.3897/popcon.7.e97064>

Аннотация

Большие данные предлагают исследователям новый источник информации о демографическом поведении населения. Одним из таких источников являются тексты пользователей социальных сетей по тем или иным демографическим вопросам. Данное исследование основывается на методах автоматического извлечения мнений пользователей социальной сети «ВКонтакте» и их последующей классификации с использованием нейросетевой модели Conversational RuBERT для изучения мнений в области репродуктивного поведения населения. Классификация высказываний проводилась с целью решения двух последовательных задач. Во-первых, выявления наличия аргументации в пользовательском комментарии, во-вторых, при наличии аргумента, определения его типа в разрезе дихотомии «личное — общественное». В процессе решения задач поиска аргументов и определения их типов было проведено шесть экспериментов, различающихся набором данных и числом классов. Примененный метод автоматического извлечения и классификации мнений пользователей социальной сети «ВКонтакте» позволил убедиться в том, что мы можем достаточно точно классифицировать подаваемые на вход пользовательские комментарии на предмет наличия довода и определения его типа в дихотомии «личное — общественное» в целях выявления представленности личных и общественных установок, ценностей, историй и мнений в изучении репродуктивного поведения.

Ключевые слова

репродуктивное поведение, дихотомия «личное — общественное», автоматическое извлечение мнений, аргументация, ВКонтакте, Conversational RuBERT

Коды JEL: C8, J1

Введение

Большие данные предлагают исследователям новый источник информации о демографическом поведении населения. Одним из таких источников являются высказывания пользователей социальных сетей по тем или иным демографическим вопросам. Изучая репродуктивное поведение населения, мы начали исследовать эти данные с определения демографической температуры первого типа (эмоционального фона специальных демографических групп пронаталистского и антинаталистского толка) [Kalabikhina et al., 2021a], затем перешли к определению демографической температуры второго типа (эмоционального фона высказываний по темам в области репродуктивного поведения и оценки демографической политики) [Калабихина и др., 2021; 2022]. Мы определяем позитивное или негативное отношение к репродуктивным аспектам (многодетность, бездетность, аборт и пр.), но пока не можем автоматически вычленять аргументы такого отношения.

Хотя на сегодняшний день методы обработки естественного языка развиваются и совершенствуются ускоренными темпами, до сих пор остаются задачи, которые не имеют однозначного и простого решения. Одной из таких задач является определение и воспроизводство аргументации автоматическими методами. Однако возможность анализа высказываний, содержащих в себе утверждение и его обоснование, представляет огромный интерес, в особенности для социальных наук. Решение этой задачи позволило бы понять не только то, о чем люди считают нужным высказать свое мнение, но и на чем это мнение основывается. В рамках настоящей работы мы предпринимаем попытку провести такой анализ для аргументации по вопросам репродуктивного поведения на основании комментариев пользователей социальной сети «ВКонтакте».

Исследование основывается на мнениях пользователей социальной сети «ВКонтакте», полученных методом автоматического извлечения, и на их последующей классификации с использованием нейросетевой модели Conversational RuBERT. Классификация мнений проводилась с целью решения двух последовательных задач. Во-первых, выявления наличия аргументации в пользовательском комментарии, во-вторых, при наличии аргумента, определения его типа в разрезе «личное — общественное».

Мы используем такой разрез, опираясь на традиционный учет уровня установок населения в контексте демографического поведения для прогнозирования поведения и принятия решений в области демографической политики. Например, мы изучаем как установки на число детей, которое человек планирует иметь, так и его/ее представления об идеальном числе детей в данное время в данном обществе.

Для решения этих задач было проведено шесть экспериментов, различающихся набором данных и числом классов (два или три). В работе приводится анализ основных метрик качества обученных моделей классификации (Accuracy и F-score) и делается вывод относительно выбора наиболее релевантных моделей для обеих поставленных задач.

Обзор литературы

В последние годы в качестве инструмента изучения общественного мнения исследователи активно используют комментарии из социальных сетей, например, таких как Reddit и Twitter. В частности, данные из социальных сетей используются регулярно для оценки эмоционального спектра [Thelwall, Stuart, 2019; Mittos et al., 2020; Liu et al., 2021; Al-Rawi et al., 2021].

Стоит выделить отдельные исследования, направленные на извлечение аргументов из массива текстов для разного рода целей или областей исследования с возможностью использования нескольких систем машинного перевода. Так, отечественные исследователи создали

аргументно-аннотированный корпус русского языка и оценили производительность различных классификаторов для анализа аргументации [Fishcheva, Kotelnikov, 2019]. В частности, метод и результаты этого исследования были применены для возможности анализа аргументов на наборе данных COVID-19 [Kotelnikov et al., 2022].

Продолжая тему извлечения аргументации, необходимо упомянуть ряд научных работ, в которых отмечается, что комментарии могут являться источником для извлечения аргументации, в частности ряд авторов проводил такие исследования на основе корпуса комментариев под видео на платформе Youtube [Sagredos, Nikolova, 2022] и под новостными постами [Ehret, Taboada, 2020]. Однако Castellano Parra с соавторами при анализе комментариев под новостными постами на порталах, а также комментариев в социальных сетях на страницах ряда испанских газет пришли к следующему выводу: как уровень вовлеченности (количество комментариев на 1 читателя), так и доля аргументированных комментариев в социальных сетях выше, чем в комментариях на новостных порталах [Castellano Parra et al., 2020]. При этом исследователи отмечают, что на обеих платформах аргументы и точки зрения принадлежат ограниченному кругу активных пользователей, что способствует монополизации дискурса отдельными лицами и группами [Jensen, 2016].

Анализ аргументов из комментариев социальных сетей набрал популярность в научно-исследовательской среде в последние годы в связи с ростом дискуссионности отдельных тематик по COVID-19: вакцинация [Melton et al., 2021; Wawrzuta et al., 2021; 2022], карантин [Karami, Anderson, 2020]. Кроме этого, комментарии используются для мониторинга и принятия оперативных решений в вопросах COVID и его последствий [Han et al., 2020; Oyeboode et al., 2021].

В научно-исследовательской среде можно встретить отдельные исследования на основе анализа комментариев, в которых выражается отношение к абортам [Hasan, Ng, 2013; Graells-Garrido et al., 2019], в частности по легализации абортотворения [Misra et al., 2017]. Стоит также отметить работы по анализу контента комментариев по другим темам, связанным с демографией: аспекты родительства [Mencarini et al., 2017], проблемы здравоохранения [Shah et al., 2019], влияние на демографические процессы различных факторов, например природных катастроф [Mandel et al., 2012] или пандемии COVID-19 (включая вакцинацию для детей) [Miao et al., 2020; Glandt, 2021; Liu, Liu, 2021; Thorpe Huerta et al., 2021; Abosedra et al., 2021], сексуальные домогательства или насилие [Andalibi et al., 2016; Xue et al., 2019; Al-Rawi et al., 2021; Lin et al., 2022], отношение к генетическим тестам [Mittos et al., 2020] и т.д.

Анализ комментариев социальных сетей в рамках демографических тематик пока недостаточно популярен в отечественной литературе. Однако применение пользовательского контента в целом в демографических исследованиях в последние годы возросло. Была произведена попытка анализа настроений русскоязычных комментариев в социальных сетях для определения «демографической температуры» в пронаталистских и антинаталистских группах пользователей социальных сетей [Kalabikhina et al., 2021a], а также анализ мнений пользователей социальных сетей по вопросам репродуктивного поведения на основе комментариев ВКонтакте [Калобихина и др., 2021; 2022]. В частности, было доказано, что комментарии могут быть использованы для анализа настроения высказываний лучше, чем тексты постов [Kalabikhina et al., 2021a].

Кроме этого, были опубликованы несколько датасетов с комментариями пронаталистов и антинаталистов, которые легли в основу данных исследований [Kalabikhina, Banin, 2020; 2021].

Другие исследователи в рамках анализа контента социальных сетей концентрируются преимущественно на оценке эмоционального фона постов [Donchenko et al., 2017] и комментариев [Sidorov, Slatnikov, 2021; Smetanin, Komarov, 2021], не предпринимая попыток извлечь аргументацию позиций пользователей.

Методологический подход к типологизации аргументов (доводов)

Особенный интерес для нас представляло не просто выявление наличия аргументации в комментариях, но и определение ее природы. Автоматизация всего спектра доводов/аргументов по различным проблемам (даже объединенным демографической тематикой) является чрезвычайно сложным делом. Однако переход к ручной обработке текстов после автоматического выделения комментариев с аргументами приводит к большим затратам времени. Поэтому первым шагом на пути автоматизации извлечения аргументов стал поиск укрупненных классов аргументов, которые могут быть использованы в обсуждении различных демографических тем.

Классификацию аргументов по типу мы рассматривали в разрезе «личное — общественное» в зависимости от того, руководствуется ли комментатор своим личным опытом или приводит в качестве аргумента некоторые общественные представления.

Такой подход основан на опыте разделения мнения в оценке персональных и идеальных (общественно поощряемых) аспектов репродуктивного поведения. Например, при оценке репродуктивных установок традиционный блок вопросов содержит вопросы об ожидаемом и идеальном числе детей. Ожидаемое число детей — число детей, которое респондент собирается иметь лично с учетом всех персональных обстоятельств. Идеальное число детей — представления респондента о норме в данное время в данном обществе. Иногда вопрос ставится о желаемом числе детей, что близко к нормативному вопросу, хотя он и адресован личным планам без учета обстоятельств.

Для данной темы — оценки репродуктивного поведения или отношения людей к мерам демографической или семейной политики — такой подход также имеет смысл. Если можно научиться выделять личные аргументы/истории и высказывания об общественных нормах, правилах демографического поведения, побуждения к действиям других людей, то мониторинг «личного» и «общественного» будет полезен для реагирования на тревожные ситуации по демографическим темам. В период, когда доля «личного» растет, необходимо более внимательно относиться к комментариям пользователей соцсетей. Поскольку соцсети (большинство из них) скорее дают нам критическую информацию по теме. Фильтрация личных аргументов сужает объем материала для обработки и выделения конкретных доводов.

Личные и общественные аргументы могут быть не только в области репродуктивной тематики. Матримониальное, миграционное, самосохранительное поведение (установки, отношения, действия) также может быть представлено личным или общественным опытом. Аргумент об отказе от вакцинации по личным обстоятельствам или аргумент об отказе от вакцинации, который обсуждается в обществе, — пример такого разделения.

Похожий подход был обнаружен в работе [Kiesel et al., 2022]. Авторы поставили цель автоматизировать поиск ценностей (используя наборы ценностей в различных социологических опросах о ценностях — Шварца, Всемирного обследования ценностей и др.). Основная цель работы — связать высказывания людей (естественный язык) с ценностными ориентациями и автоматизировать эту работу, обучив искусственный интеллект этим связям. Интересно для нашего исследования, что описываемая система ценностей, используемая авторами, содержала деление ценностей на личные и общественные.

Данные

В данном исследовании используются два набора демографических данных. Основным датасетом является набор репродуктивных данных, которые представляют собой комментарии пользователей сети «ВКонтакте» по тематике, связанной с рождением детей [Kalabikhina,

Banin, 2020; 2021]. Дополнительно используемый датасет представляет собой описанный в работе [Kotelnikov et al., 2022] набор высказываний пользователей «ВКонтакте» по разным аспектам коронавирусной инфекции.

Репродуктивные данные

Сбор данных

Источник текстовых данных по репродуктивной тематике — тематические группы в социальной сети «ВКонтакте» (vk.com). На первом этапе обработки с помощью встроенного API (application programming interface) по ключевым словам («мама», «мамочки», «дети», «ребенок», «малыш», «здоровье», «рождение», «беременность», «родители», «чайлдфри», «чайлдхейт» и т.д.) были собраны уникальные адреса тематических групп в виде *vk.com/<уникальный идентификатор группы>*, которые, исходя из типа ключевого слова, условно делились на две подгруппы: антинаталисты (ключевые слова: «чайлдфри», «чайлдхейт» и т.п.) и пронаталисты (ключевые слова: «семья», «дети», «малыш», «беременность»). Было собрано около 1000–1500 уникальных адресов групп с данными о количестве участников.

На втором этапе из выборки были исключены группы, связанные с рекламой, а также группы с малой активностью участников. Поскольку во всех группах встречались рекламные посты, то предпочтение отдавалось группам, в которых присутствовали содержательные тексты на тему материнства и отцовства, демографической политики и т.п. Для отбора использовался автоматизированный метод на основе отсеечения наиболее неактивных групп (небольшое количество подписчиков, низкая публикационная активность), отсеечения групп, для которых было характерно редкое упоминание ключевых слов, низкая активность в комментариях под постами. Пороговые параметры отсеечения определялись итерационно. Для групп пронаталистов порог отсеечения по количеству пользователей соответствовал 10 000 подписчиков, это позволило собрать 341 целевую группу пронаталистов. Количество групп антинаталистов существенно меньше, но активность в таких группахкратно больше, поэтому фильтр должен учитывать подобную специфику. Для антинаталистов порог отсеечения соответствовал пятистам участникам (подобные группы малочисленны по составу), что позволило отобрать восемь активных групп антинаталистов.

- Пронаталисты (максимальное количество подписчиков 1 482 303, минимальное — 72 570, в среднем 309 тысяч подписчиков на группу).
- Антинаталисты (максимальное количество подписчиков 61 071, минимальное — 619, в среднем 8950 подписчиков на группу).

После формирования окончательного списка групп осуществлялся сбор текстовой информации (посты и комментарии к ним) из групп. На основании собираемой информации формировался корпус комментариев: все слова приводились к нижнему регистру, стоп-слова¹ удалялись при помощи функций из библиотеки nltk или gensim, удалялась пунктуация, исключались числовые данные. Для уменьшения объема текстовых данных дополнительно проводились стемминг (удаление окончаний слова) или лемматизация (приведение слова к начальной форме при помощи лемматизатора MyStem). Структура выборки и список основных групп представлены в [Kalabikhina, Banin, 2020; Kalabikhina, Banin, 2021]. После сбора текстовой информации проводился поиск по ключевым словам в собранных текстах для определения наиболее релевантных текстов для дальнейшей разметки. Таблица 1 содержит список тем и ключевых слов, которые использовались для отбора текстов.

¹ Стоп-слова — это часто используемые слова, которые не вносят никакой дополнительной информации в текст. Например, мы удаляли союзы, местоимения и предлоги, не несущие смысловой нагрузки.

Таблица 1. Списки ключевых слов для извлечения высказываний по репродуктивной тематике

Тема	Характерные слова
Материнский капитал/пособие	маткапитал, материнский капитал, выплаты, пособие
Аборт	аборт
Многодетность	многодетный, многодетность, много детей
Бездетность	childfree, чайлдфри, бездетный, нет детей, без детей
Родительские отпуска	декрет, отпуск
Индивидуализм	в свое, эгоист, ответственность, для себя, личность, развиваться

Источник: составлено авторами.

Разметка репродуктивных данных по позиции автора и аргументации

На первом этапе полученные высказывания пользователей по репродуктивной тематике размечались в соответствии с позицией автора по отношению к заданной теме («за», «против», «другое»). В рамках данной разметки были отброшены высказывания, нерелевантные к заданным темам. На втором этапе в размеченные по авторской позиции данные была добавлена разметка по наличию в них аргументации.

Для разметки по авторской позиции предложения из сформированной выборки случайным образом отбирались для разметки аннотаторами. Каждое предложение размечалось тремя аннотаторами. Поскольку в каждом предложении могло обсуждаться несколько вопросов, то аннотатор каждое предложение размечал по всем шести темам. Оценка предложения выводилась на основании оценок нескольких аннотаторов посредством голосования, по большинству голосов. Таблица 2 содержит статистику полученного репродуктивного датасета по темам и по позициям [Kalabikhina, Banin, 2021]. Суммарный объем размеченной выборки составил 5412 комментариев.

Таблица 2. Распределение авторских оценок по темам, относящимся к рождению детей

Тема	Релевантно	За	Против	Прочее
Аборты	1374	709	161	504
Многодетность	341	75	153	113
Родительские отпуска	992	201	376	435
Индивидуализм	739	464	119	156
Маткапитал/Пособия	813	184	370	259
Бездетность	1422	853	134	435
Всего	5681*	2486	1313	1902

Источник: составлено авторами.

Примечание: общее число оценок (*) в таблице больше, чем общее число анализируемых комментариев, так как некоторые комментарии содержали мнения сразу по нескольким темам.

На втором этапе созданный датасет по репродуктивной тематике был размечен по наличию в высказывании аргументации. Аннотаторы размечали, есть ли в высказывании аргумент, т.е. может ли высказывание использоваться как довод в споре. При наличии аргументации высказывания дополнительно размечались по типу аргументации: общественное или личное.

Предварительный анализ комментариев

До этапа автоматического извлечения доводов был проведен предварительный ручной анализ комментариев из выгрузки с целью поиска возможных подходов к автоматической классификации доводов. Примеры комментариев по каждой из анализируемых тем представлены в таблице 3 (орфография и пунктуация авторов сохранены).

Таблица 3. Примеры комментариев из анализируемой базы по темам интереса

Тема	Примеры комментариев
Аборт	<ul style="list-style-type: none"> • «в 20–30-е годы возрождалась промышленность быстрыми темпами, да было нелегко, но советские люди войну выстояли, а они тут про аборты как фактор сокращения численности населения, бред»; • «и как же обойти? уехать в другую страну делать аборт? не у всех есть такая возможно»
материнский капитал/пособия	<ul style="list-style-type: none"> • «материнский капитал не мотивирует от него нет толку»; • «маткапитал служит хорошую службу тем, у кого есть уже свой угол, квартира +одна комната, бездомным (ой, снимающим) рассчитывать там не на что»
бездетность	<ul style="list-style-type: none"> • «нет детей нет проблем»; • «Без детей все ссоры в семье разгораются.»
многодетность	<ul style="list-style-type: none"> • «зачем много детей, считаю одного надо одеть обути вырастить накормить напоить и образование дать»; • «Много детей = страшная мать.»
отпуск	<ul style="list-style-type: none"> • «недавно была на собеседовании, hr-менеджер девушка примерно моя ровесница, в конце собеседования задала эпичный вопрос, когда я планирую декретный отпуск»; • «ведь декретницы сидят не у вас на шее и не едят с вашего кармана»

Примечание: тексты комментариев приведены в оригинальном виде.

Источник: составлено авторами.

Видно, что не все комментарии содержат аргумент. Кроме того, отмечается большое число различных типов содержательных аргументов по каждой теме, что затрудняет выбор унифицированных аргументов для использования их в разных темах в целях автоматической выгрузки таких аргументов. Мы пытались найти хотя бы несколько типовых доводов для всех наших тем — репродуктивное поведение и отношение к пронаталистской демографической политике. По мнению авторов статьи, разделение доводов на позитивные и негативные (в отличие от измерения эмоционального фона собственно текстов) мало что дает в практическом смысле, поскольку такая классификация частично дублирует оценку эмоционального фона самих текстов и доводов в них. Принять политическое решение в зависимости от изменения доли позитивных/негативных доводов невозможно без дополнительного ручного анализа содержания. В связи с этим было принято решение использовать разделение аргументов на более

универсальные типы — личные и общественные, то есть высказывания о личных переживаниях, установках, ценностях, историях или высказывания об идеальных установках, ценностях, нормах, о том, «как надо» себя вести.

Классификации доводов по критериям «личные» и «общественные»

В демографии такой критерий имеет значение для определения демографического поведения. Например, в опросах о числе детей имеется несколько типов вопросов, которые нацелены различать «общественные» установки и личные действия. 1) Какое идеальное число детей, по Вашему мнению? 2) Сколько детей именно Вы планируете иметь в соответствии с Вашими возможностями? Показано на многих датасетах по разным странам, что ответ на второй вопрос очень хорошо предсказывает уровень рождаемости соответствующих поколений.

Что нам даст мониторинг личных и общественных доводов? Мы по-прежнему не сможем принять политическое решение в зависимости от изменения доли личных/общественных доводов без дополнительного ручного анализа содержания. Но для нас это будет новый способ измерения ситуации: увеличение доли личных доводов — сигнал об обострении ситуации, сигнал для внимательного изучения текстов доводов в ручном режиме в этот период времени. Если от рассуждений «вообще» люди переходят к рассказу о себе, как правило, это сопряжено с ростом проблем и в наших темах.

Для тем про многодетность, бездетность, аборт, пособия, отпуска также выбираются личные и общественные доводы.

Индивидуальные случаи посвящены либо рассказам о том, что произошло/происходит с автором довода или его семьей, либо о том, что он(а) думает о своей личной ситуации.

Общественные случаи — размышления авторов о том, как (не) надо жить, вести себя в демографическом контексте всем жителям страны/региона/социальной группы.

Примеры комментариев с доводами по категориям «личное» и «общественное» представлены в таблице 4.

Таблица 4. Примеры комментариев с доводами по категориям «личное» и «общественное»

Тип довода (личное / общественное)	Комментарий
Личное	<ul style="list-style-type: none"> «я замужем 5 лет, детей нет! неоднократно меняла работу, и никто не интересовался почему у нас нет детей и когда пойду в декрет! только опыт работы, образование, характеристики и личные качества»
Личное	<ul style="list-style-type: none"> «А еще я знаю, что это стыдно ощущать себя нищим, что в тебя тыкают пальцем твои сверстники и говорят, что ты из многодетной семьи нищоброд, когда у тебя нет своего угла дома, когда тетради покупают самые дешевые и их не хватает, поэтому я могу не по наслышан сказать, что дети- это ответственность полностью родителей и нужно думать на что ты их будешь содержать, сможешь ли ты обеспечить им не только будущее, но и просто нормальное существование.»
Личное	<ul style="list-style-type: none"> «Лично мои Дети, это мой смысл жизни, я их очень люблю, и люблю больше жизни!»
Личное	<ul style="list-style-type: none"> «У меня трое детей тоже беспорядок наводят лучше так чем без детишек.без детей нет смысла жизни»

Тип довода (личное / общественное)	Комментарий
Личное	• «Моя мама родила 12 детей и от неё никогда не пахло плохо и из ничего она делала шедевры.....девочки давайте не будем забывать про себя,как бы трудно нам не было, по одной простой причине, что мы женщины, а ещё для того, что — бы нами гордились наши дети и наши мужья»
Личное	• «я и без калькулятора прекрасно могу понять, что не потяну содержать жену с ребенком в декрете: цены на детское всё видел пару раз — мой внутренний еврей мне подробно сказал что он заранее обо мне думает и куда мне идти, если я вдруг решу сделать свою плохонькую ксерокопию»
Общественное	• «в странах с запретом абортотв зашкаливает уровень преступности»
Общественное	• «автор правильно сказали — никаких детей пока на ноги не встанете, а еще лучше пока для себя в полной мере не поживете»
Общественное	• «общество получает от чайлдфри больше плюшек чем от детных, у нас нет декрета, отпрашиваний с работы из-за проблем и болезней детей, мы не занимаем места в школах, садах и поликлиниках, на всякие выплаты по уходу, мат»
Общественное	• «а потом детей обоих в детдом сдать?что тогда будет?потратили же уже маткапитал на улучшение жилищных условий например»
Общественное	• «это ненормальная ситуация в стране, когда эп нищенские, цены на все конские, детское пособие тоже гроши, а женщина в декрете выпадает из жизни и становится мега-зависимой»
Общественное	• «так что я лично за аборты, а не за испорченные жизни»

Примечание: тексты комментариев приведены в оригинальном виде.

Источник: составлено авторами.

Следует отметить, что в ходе разметки встречались и спорные случаи. Некоторые комментарии было сложно отнести к одной из двух категорий: личное или общественное. Например, отдельные комментарии содержали в себе оба элемента одновременно: «Вот я воспитываю троих детей на пособие и все остальные также должны!». В данном предложении, с одной стороны, описан личный опыт автора, с другой стороны, представления о том, как должно быть устроено общество. Данный комментарий стоит выделить также потому, что он, вероятно, содержит элемент иронии. Автор, скорее всего, цитирует своих оппонентов, высмеивая их мнение. Комментариев с иронией и цитированием в выгрузке было достаточно много, однако на данном этапе они не исключались, так как с точки зрения текущих исследовательских задач это не требуется.

Ручная разметка комментариев

Ручная разметка комментариев выполнена шестью аннотаторами (авторами этой статьи). Каждый комментарий оценивался тремя аннотаторами независимо друг от друга. Разметка комментариев происходила по двум параметрам: 1) наличие довода (есть довод / нет довода); 2) тип довода (личное / общественное). На первом этапе аннотатор должен был определить, содержится ли в комментарии довод. Доводы — аргументационные высказывания, которые

могут быть использованы для убеждения оппонента относительно определенной точки зрения. Доводы могут быть в поддержку или в опровержение точки зрения. В нашем случае нам не важно — довод «за» или «против». Высказывание оценивалось как довод, если оно содержит утверждение, которое может использоваться как аргумент в споре, в убеждении оппонента.

Пример 1 (нет довода, хотя и есть негативное отношение к абортам):

«сами зародыши и эмбрионы, их, конечно не заботят: — они никогда не выступают против селективных абортов девочек (кавказ, средняя азия, арабские страны, китاي и т»

Пример 2 (нет довода, хотя и есть позитивное отношение к декретному отпуску):

«то, что она сидит дома в декрете — не значит, что ей (...хорошо...) и она отдыхает»

Пример 3 (есть довод, негативное отношение к декретному отпуску, потому что отпуск приводит к потере профессиональных навыков у женщины):

«она выйдет из декрета, придёт на работу, там выяснится, что работать она не умеет — зачѣт в работе не нужен, нужен результат»

На втором этапе выполнялась разметка комментариев с доводами (определены в ходе первого этапа разметки): аннотатор оценивал, относится ли данный довод к категории «личное» или «общественное». Таким образом, каждый комментарий репродуктивного датасета содержит результаты независимой разметки трех аннотаторов.

В среднем доля совпадающих оценок о наличии довода и его типе составляла более 70%, что считается удовлетворительным результатом (см. таблицы 5 и 6).

Таблица 5. Результаты ручной разметки базы комментариев. Число совпадений по параметру наличия довода между двумя аннотаторами

Аннотаторы	Число совпадений: довод / не довод					
	Аннотаторы 1 и 3	Аннотаторы 1 и 2	Аннотаторы 2 и 3	Аннотаторы 4 и 5	Аннотаторы 4 и 6	Аннотаторы 5 и 6
% совпадений	86,4	84,2	78,6	73,8	73,3	71,2
Число совпадений	2336	2274	2124	1997	1985	1928
Всего размечено комментариев	2705	2705	2705	2707	2707	2707

Источник: составлено авторами.

Таблица 6. Результаты ручной разметки базы комментариев. Число совпадений по параметру типа довода между двумя аннотаторами

Аннотаторы	Число совпадений: личное / общ.					
	Аннотаторы 1 и 3	Аннотаторы 1 и 2	Аннотаторы 2 и 3	Аннотаторы 4 и 5	Аннотаторы 4 и 6	Аннотаторы 5 и 6
% совпадений	82,1	90,3	77,8	84,0	74,5	80,6
Число совпадений	724	646	526	516	533	625
Всего комментариев (которые оба разметчика отметили как содержащие довод)	883	719	677	617	717	777

Примечание: учитываются только случаи, когда оба аннотатора отметили комментарий как содержащий довод.

Источник: составлено авторами.

Описательная статистика результатов разметки каждого аннотатора представлена в таблице 7.

Таблица 7. Описательная статистика результатов разметки каждого аннотатора

Аннотатор	Доводы		из них: Личные истории		Всего размечено комментариев
	абс.	%	абс.	%	
1-я группа комментариев					
Аннотатор 1	1034	38,2	381	36,8	2705
Аннотатор 2	835	30,9	303	36,3	2705
Аннотатор 3	1101	40,7	578	52,5	2705
2-я группа комментариев					
Аннотатор 4	953	35,2	335	35,2	2707
Аннотатор 5	1051	38,8	323	30,7	2707
Аннотатор 6	1280	47,3	516	40,3	2707

Источник: составлено авторами.

Данные COVID-19

Данные COVID-19, используемые в нашем исследовании, представляют собой комментарии пользователей сети «ВКонтакте» по поводу новостных сообщений тематики COVID-19. Собранные комментарии были отобраны по ключевым словам, соответствующим трем аспектам, связанным с пандемией COVID-19, а именно «маски», «карантин» и «вакцинация» [Nugamanov et al., 2021].

На первом этапе комментарии были размечены по позиции автора к заданным аспектам («за», «против», «другое»). На втором этапе размеченные высказывания были дополнительно аннотированы по наличию аргумента, обосновывающего позицию, и к какому конкретному аспекту [Kotelnikov et al., 2022]: наличие аргумента «за», наличие аргумента «против», нет аргумента. Всего было размечено 9550 высказываний, из них около 2000 содержали аргументы. В данном исследовании разметка по наличию аргумента независимо от темы используется как датасет для предобучения модели Conversational RuBERT. Добавляя данные COVID-19 к репродуктивному датасету, авторы планировали исследовать возможность улучшения результатов обработки целевой коллекции за счет предобучения на данных другой тематики, но при этом также относящихся к широкой области изучения народонаселения и демографии.

Эксперименты

В данном исследовании решались две задачи классификации: по наличию¹ и по типу аргумента (общественное или личное). В общей сложности было протестировано шесть экспериментов, различающихся тем, какую задачу они решают (эксперименты 1–3 — наличие аргумента, эксперименты 4–6 — тип аргумента) и какие данные в них использовались (в экспериментах 2 и 3

¹ Наличие аргумента мы сначала пытались искать, используя союзные слова («потому что», «так как» и пр.). Однако, оценив результаты, остановились на методе разметки комментариев по наличию аргумента для машинного обучения впоследствии.

использовались репродуктивные и данные COVID-19, в остальных — только репродуктивные). Во всех экспериментах размеры тренировочной, тестовой и валидационной выборки были взяты в соотношении примерно 80:10:10. Разделение комментариев по данным группам в таком соотношении было произведено так, чтобы выборки были равномерно распределены по основным интересующим нас темам (см. раздел «Сбор данных»).

Эмпирическая часть исследования проводилась в среде Python с использованием PyTorch и библиотек Transformers и Scikit Learn. Все эксперименты основывались на модели Conversational RuBERT¹, представляющей собой версию предобученной русскоязычной языковой модели RuBERT, дополнительно обученную на данных социальных сетей и пользовательских диалогах. В экспериментах использовались следующие параметры: learning rate (коэффициент скорости обучения модели) = 0,0005, batch size (число объектов тренировочной выборки, представленных в одном обучающем сете для одной итерации) = 64. Выбор числа эпох (число прохождений тренировочных данных через модель с последующим обновлением весов модели) от 4 до 6 основывался на лучшем показателе F-score на валидационном множестве. Для оценки качества модели использовались параметры Accuracy и F1-score (в сокращенном виде приведены в таблицах 8 и 9, более подробная информация о качестве работы модели — в Приложении в таблицах П1–П6).

Результаты

Определение наличия аргумента

В рамках первой задачи по определению наличия аргумента комментарии пользователей «ВКонтакте» были разделены на два класса: «1» — есть аргумент, то есть не менее двух из трех аннотаторов согласились, что автор комментария приводит некую аргументацию, «0» — аргумента нет, то есть аргументацию в комментарии отметил лишь один из трех аннотаторов или не отметил никто. Для решения этой задачи рассматривалось три варианта эксперимента (метрики для сравнения представлены в таблице 8). Во всех экспериментах соотношение числа комментариев по классам в среднем составляло 40:60 для комментариев, содержащих аргумент, и комментариев, где аргументация отсутствует.

В эксперименте 1 обучение классификационной модели проводилось на репродуктивных данных — комментариях пользователей «ВКонтакте» на тему репродуктивного поведения и оценки мер демографической политики. Суммарный объем тренировочной, тестовой и валидационной выборки составил 5410 комментариев, предварительно обработанных по стандартному алгоритму: удалены заглавные буквы, пунктуация, стоп-слова, пустые комментарии.

В эксперименте 2 к репродуктивным данным были дополнительно добавлены комментарии пользователей «ВКонтакте» по темам, связанным с пандемией коронавируса (вакцинация, ношение масок, карантинные ограничения и т.д.). Все новые комментарии были включены только в тренировочную выборку. Суммарный объем дополнительных данных составил 6716 комментариев, представляющих собой обучающую выборку данных COVID-19.

В эксперименте 3 обучение проводилось в два этапа: предварительное обучение на данных COVID-19, далее дообучение и оценка качества модели на репродуктивных данных.

Представленные в таблице 8 результаты показывают, что применением дополнительных данных COVID-19 на текущий момент не удалось улучшить результаты выделения аргументов на целевой репродуктивной коллекции: лучшие результаты показала модель, которая обучена на тренировочной коллекции репродуктивных данных.

1 <https://huggingface.co/DeepPavlov/rubert-base-cased-conversational>

Таблица 8. Модель классификации по наличию аргумента («0» — аргумента нет, «1» — аргумент есть)

	F1-score для класса «0»	F1-score для класса «1»	Accuracy
Эксперимент 1: только репродуктивные данные	0,81	0,61	0,75
Эксперимент 2: репродуктивные и данные COVID-19	0,82	0,49	0,73
Эксперимент 3: обучение на данных COVID-19, дообучение на репродуктивных	0,79	0,48	0,70

Примечания: в эксперименте 1 использованы только репродуктивные данные, в эксперименте 2 репродуктивные и данные COVID-19 использованы в тренировочной выборке, тогда как в тестовой и валидационной только демографические данные, в эксперименте 3 данные COVID-19 использовались для предобучения модели, тогда как последующее обучение и оценка проводились на репродуктивных данных.

Источник: составлено авторами.

Определение типа аргумента

Особенный интерес для нас представляло не просто выявление наличия аргументации в комментариях, но и определение содержания аргументов, хотя бы на высоком уровне обобщения, когда классификация типов аргументов не чувствительна к теме аргументации, иначе это привело бы к снижению качества модели.

Как описано выше, классификация аргументов по типу рассматривалась в разрезе «личное — общественное» в зависимости от того, руководствуется ли комментатор своим личным опытом или приводит в качестве аргумента некоторые общественные представления.

Эта задача оказалась несколько сложнее, чем просто определение наличия аргументации, из-за большого числа спорных случаев, когда высказывания содержали элементы обоих типов, но аннотаторам нужно было выбрать только один класс для каждого комментария.

Для решения этой задачи было решено провести эксперименты двух типов: классификация по двум классам на выборке, содержащей только комментарии с аргументацией, и классификация по трем классам на полной выборке (метрики качества модели представлены в таблице 9).

В экспериментах 4 и 5 общий объем репродуктивных данных для обучения был сокращен более чем в два раза — до 2049 комментариев, включающих только те, в которых не менее двух из трех аннотаторов отметили наличие аргумента. В эксперименте 5 класс «1» содержал комментарии с личной аргументацией по мнению не менее двух из трех аннотаторов, класс «0» — все остальные (включая спорные случаи). Соотношение числа наблюдений класса «0» к числу наблюдений класса «1» составило примерно 65:35. Аналогично в эксперименте 5 такое же разделение было проведено для аргументации по типу общественного. Соотношение числа наблюдений класса «0» к числу наблюдений класса «1» составило примерно 45:55.

В эксперименте 6 модель была обучена на полном множестве репродуктивных данных, разделенных на три класса: «0» — аргументация отсутствует, «1» — аргументация общественного типа, «2» — личная аргументация. Число комментариев по трем классам «аргумент отсутствует» — «общественный аргумент» — «личный аргумент» составило примерно 70:15:15.

Таблица 9. Модель классификации по типу аргумента («личный» — на основе личного опыта, «общественный» — на основе общественных представлений)

	F1-score для класса «0»	F1-score для класса «1»	F1-score для класса «2»	Accuracy
Эксперимент 4: два класса, акцент на личные доводы	0,86	0,71	-	0,81
Эксперимент 5: два класса, акцент на общественные доводы	0,70	0,81	-	0,77
Эксперимент 6: три класса	0,78	0,52	0,34	0,67

Примечания: в экспериментах 4 и 5 используются репродуктивные данные в усеченном виде, отобранные по принципу «не менее 2 из 3 аннотаторов согласились, что в комментарии есть аргументация». В эксперименте 4 класс «1» содержит комментарии с личными доводами по мнению не менее 2 из 3 аннотаторов, все остальные комментарии относятся к классу «0». В эксперименте 5 класс «1» содержит комментарии с общественными доводами по мнению не менее 2 из 3 аннотаторов, все остальные комментарии относятся к классу «0». В эксперименте 6 репродуктивные данные использованы в полном объеме и разделены на три класса («0» — нет аргумента, «1» — общественный аргумент, «2» — личный аргумент).

Источник: составлено авторами.

Лучшие результаты классификации по определению наличия аргумента среди первых трех экспериментов показали эксперимент 1 и эксперимент 2. Вероятное объяснение может заключаться в том, что в эксперименте 1 в отличие от трех других модель обучалась только на репродуктивных данных, вследствие чего такая модель была меньше подвержена риску некорректной классификации по причине различий в специфике тем обсуждения (репродуктивное поведение и пандемия). В то же время эксперимент 2 отличался наибольшим количеством наблюдений в тренировочной выборке, что является логичным объяснением ее более высокого качества по сравнению с другими моделями.

В рамках решения второй задачи по определению типа аргумента на основании сравнения метрика качества эксперименты 4 и 5 оказались ожидаемо лучше эксперимента 6, так как, во-первых, решали более простую задачу (два класса вместо трех), а во-вторых, содержали более чистые данные, потому что при сокращении выборки автоматически были исключены многие спорные случаи. К тому же соотношение наблюдений по классам в экспериментах 4 и 5 было более равномерным, чем в эксперименте 6.

Заключение

При аннотировании комментариев пользователей социальной сети «ВКонтакте» почти в 40% выбранных нами комментариев по темам, связанным с репродуктивным поведением, было обнаружено наличие довода (аргумента в споре за высказанное автором негативное или позитивное отношение к вопросу по определенной демографической тематике). Из высказываний, имеющих аргумент, почти 40% апеллировали к личной истории, остальные были причислены к общественным аргументам (общественные установки, ценности, нормы). Эти результаты позволяют нам сделать вывод о наличии достаточной базы для анализа аргументов в пользу той или иной «демографической» точки зрения пользователей сети, а также для вычленения доводов личного и общественного типа.

Метод автоматического извлечения мнений пользователей социальной сети «ВКонтакте» и их последующей классификации с использованием нейросетевой модели Conversational RuBERT позволил нам убедиться в том, что возможно достаточно точно классифицировать подаваемые на вход пользовательские комментарии по наличию довода и его типу в дихотомии «личное — общественное».

В процессе решения задач поиска аргументов и определения их типов было проведено шесть экспериментов, различающихся набором данных и числом классов.

Таким образом, как нам представляется, в целом в рамках обеих задач модель научилась достаточно точно классифицировать подаваемые на вход пользовательские комментарии, что открывает перспективы её дальнейшего использования с целью автоматизации работы с данными такого рода. В частности, наш подход может быть использован для постоянного мониторинга данных социальных сетей в целях своевременной диагностики повышения уровня критических высказываний личного типа на демографические темы и совершенствования пронаталистской социально-демографической политики. Предположение о возможности применения разработанного алгоритма к диагностике другого типа демографического поведения для выявления масштабов высказываний личного типа и динамики доли таких высказываний ставит перед авторами исследования новые цели — апробировать данный алгоритм в анализе других типов демографического поведения (самосохранительного, матримонийального, миграционного).

Благодарности

Исследование выполнено в рамках НИР «Воспроизводство населения в социально-экономическом развитии». Работа Натальи Лукашевич по датасету COVID-2019 и его использованию в текущем исследовании по репродуктивному поведению финансировалась за счет гранта РНФ No. 21-71-30003.

Список литературы

- Калабихина И.Е., Лукашевич Н.В., Банин Е.П., Алибаева К.В., Ребрей С.М. (2021) Автоматическое извлечение мнений пользователей социальных сетей по вопросам репродуктивного поведения // Программные системы: теория и приложения 4(51): 33–63. <https://doi.org/10.25209/2079-3316-2021-12-4-33-63>
- Калабихина И.Е., Лукашевич Н.В., Банин Е.П., Алибаева К.В. (2022) Автоматический анализ репродуктивных ценностей пользователей сети ВКонтакте // Интеллектуальные системы. Теория и приложения 1(26): 90–96.
- Abosedra S., Laopodis N.T., Fakihi A. (2021) Dynamics and asymmetries between consumer sentiment and consumption in pre- and during-COVID-19 time: Evidence from the US // The Journal of Economic Asymmetries: 24: e00227. <https://doi.org/10.1016/j.jeca.2021.e00227>
- Al-Rawi A., Grepin K., Li X., Morgan R., Wenham C., Smith J. (2021) Investigating public discourses around gender and COVID-19: A social media analysis of Twitter data // Journal of Healthcare Informatics Research: 5: 249–69. <https://doi.org/10.1007/s41666-021-00102-x>
- Andalibi N., Haimson O.L., De Choudhury M., Forte A. (2016) Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In: Proceedings of the 2016 CHI conference on human factors in computing systems, San Jose (USA), May 7-12. Association for Computing Machinery, New York, 3906–18. <https://doi.org/10.1145/2858036.2858096>
- Castellano Parra O., Meso Ayerdi K., Peña Fernandez S. (2020) Behind the comments section: The ethics of digital native news discussions. Digital Native News Media: Trends and Challenges: 8(2). <https://doi.org/10.17645/mac.v8i2.2724>

- Donchenko D., Ovchar N., Sadovnikova N., Parygin D., Shabalina O., Ather D. (2017) Analysis of comments of users of social networks to assess the level of social tension // *Procedia Computer Science*: 119: 359–67. <https://doi.org/10.1016/j.procs.2017.11.195>
- Ehret K., Taboada M. (2020) Are online news comments like face-to-face conversation? A multi-dimensional analysis of an emerging register // *Register Studies*: 2(1): 1–36. <https://doi.org/10.1075/rs.19012.ehr>
- Fishcheva I., Kotelnikov E. (2019) Cross-lingual argumentation mining for Russian texts. In: W.M.P van der Aalst et al. (eds) *Analysis of Images, Social Networks and Texts*. International Conference, Kazan (Russia), July 17–19. Springer, Cham, 134–44. https://doi.org/10.1007/978-3-030-37334-4_12
- Glandt K., Khanal S., Li Y., Caragea D., Caragea C. (2021) Stance Detection in COVID-19 Tweets. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint 46 Conference on Natural Language Processing (Volume 1: Long Papers)*. Stroudsburg (USA), August 1–6. Association for Computational Linguistics, 1596–611. <https://doi.org/10.18653/v1/2021.acl-long.127>
- Graells-Garrido E., Baeza-Yates R., Lalmas M. (2019) How representative is an abortion debate on twitter? In: *Proceedings of the 10th ACM Conference on Web Science*, Boston (USA), 30 June–3 July. Association for Computing Machinery, New York, 133–4. <https://doi.org/10.1145/3292522.3326057>
- Han X., Wang J., Zhang M., Wang X. (2020) Using social media to mine and analyze public opinion related to COVID-19 in China // *International Journal of Environmental Research and Public Health*: 17(8): 2788. <https://doi.org/10.3390/ijerph17082788>
- Hasan K.S., Ng V. (2013) Stance classification of ideological debates: Data, models, features, and constraints. In: *Proceedings of the sixth international joint conference on natural language processing*, Nagoya (Japan), October 14–18. Asian Federation of Natural Language Processing, 1348–56. URL: <https://aclanthology.org/I13-1191.pdf>
- Kalabikhina I.E., Banin E.P. (2020) Database “Pro-family (pronatalist) communities in the social network VKontakte” // *Population and Economics*: 4(3): 98–103. <https://doi.org/10.3897/popecon.4.e60915>
- Kalabikhina I.E., Banin E.P. (2021) Database “Childfree (antinatalist) communities in the social network VKontakte” // *Population and Economics*: 5(2): 92–6. <https://doi.org/10.3897/popecon.5.e70786>
- Kalabikhina I.E., Banin E.P., Abduselimova I.A., Klimenko G.A., Kolotusha A.V. (2021a) The measurement of demographic temperature using the sentiment analysis of data from the social network VKontakte // *Mathematics*: 9(9): 987. <https://doi.org/10.3390/math9090987>
- Karami A., Anderson M. (2020) Social media and COVID-19: Characterizing anti-quarantine comments on Twitter // *Proceedings of the Association for Information Science and Technology*: 57(1): e349. <https://doi.org/10.1002/pra2.349>
- Kiesel J., Alshomary M., Handke N., Cai X., Wachsmuth H., Stein B. (2022) Identifying the Human Values behind Arguments. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin (Ireland), May 22–27. Association for Computational Linguistics, 4459–71. <https://doi.org/10.18653/v1/2022.acl-long.306>
- Kotelnikov E., Loukachevitch N., Nikishina I., Panchenko A. (2022) RuArg-2022: Argument Mining Evaluation. In: *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”*, Moscow (Russia), June 15–18, 333–47. <https://doi.org/10.28995/2075-7182-2022-21-333-348>
- Liu A.E., Young J.A., Guarino J.E. (2022) Mother-Daughter sexual abuse: An exploratory study of the experiences of survivors of MDSA using Reddit // *Children and Youth Services Review*: 138: 106497. <https://doi.org/10.1016/j.childyouth.2022.106497>
- Liu S., Liu J. (2021) Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis // *Vaccine*: 39(39): 5499–505. <https://doi.org/10.1016/j.vaccine.2021.08.058>

- Liu S., Li J., Liu J. (2021) Leveraging transfer learning to analyze opinions, attitudes, and behavioral intentions toward COVID-19 vaccines: social media content and temporal analysis // *Journal of Medical Internet Research*: 23(8): e30251. <https://doi.org/10.2196/30251>
- Mandel B., Culotta A., Boulahanis J., Stark D., Lewis B., Rodrigue J. (2012) A demographic analysis of online sentiment during hurricane Irene. In: *Proceedings of the 2012 Workshop on Language in Social Media*, Montréal (Canada), June 7. Association for Computational Linguistics, 27–36. URL: <https://aclanthology.org/W12-2104.pdf>
- Melton C.A., Olusanya O.A., Ammar N., Shaban-Nejad A. (2021) Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence // *Journal of Infection and Public Health*: 14(10): 1505–12. <https://doi.org/10.1016/j.jiph.2021.08.010>
- Mencarini L., Hernandez Farías D.I., Lai M., Patti V., Sulis E., Vignoli D. (2017) Happy parents' tweets: An exploration of Italian Twitter data using sentiment analysis. *Demographic Research*: 40: 25. <https://doi.org/10.4054/DemRes.2019.40.25>
- Miao L., Last M., Litvak M. (2020) Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.nlpCOVID19-2.19>
- Misra A., Oraby S., Tandon S., Ts S., Anand P., Walker M. (2017) Summarizing dialogic arguments from social media. In: V. Petukhova, Y. Tian (eds.) *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken (Germany), August 15-17, 126-36 <https://doi.org/10.21437/SemDial.2017-14>
- Mittos A., Zannettou S., Blackburn J., De Cristofaro E. (2020) Analyzing genetic testing discourse on the Web through the lens of Twitter, Reddit, and 4chan // *ACM Transactions on the Web (TWEB)*: 14(4): 1–38. <https://doi.org/10.1145/3404994>
- Nugamanov E., Loukachevitch N., Dobrov B. (2021) Extracting sentiments towards COVID-19 aspects. In: A. Pozanenko et al. (eds.) *Supplementary Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021)*, Moscow (Russia), October 26-29, 299–312. URL: [paper24.pdf](https://arxiv.org/abs/2110.00000) (ceur-ws.org)
- Oyebode O., Ndulue C., Adib A., Mulchandani D., Suruliraj B., Orji F.A., Chambers C.T., Meier S., Orji R. (2021) Health, psychosocial, and social issues emanating from the COVID-19 pandemic based on social media comments: text mining and thematic analysis approach // *JMIR medical informatics*: 9(4):e22734. <https://doi.org/10.2196/22734>
- Sagredos C., Nikolova E. (2022) 'Slut I hate you': A critical discourse analysis of gendered conflict on YouTube // *Journal of Language Aggression and Conflict*: 10(1): 169–96. <https://doi.org/10.1075/jlac.00065.sag>
- Shah Z., Martin P., Coiera E., Mandl K.D., Dunn A.G. (2019) Modeling spatiotemporal factors associated with sentiment on Twitter: synthesis and suggestions for improving the identification of localized deviations // *Journal of Medical Internet Research*: 21(5): e12881. <https://doi.org/10.2196/12881>
- Sidorov N., Slastnikov S. (2021) Some Features of Sentiment Analysis for Russian Language Posts and Comments from Social Networks // *Journal of Physics: Conference Series* 1740(1): 12–36. <https://doi.org/10.1088/1742-6596/1740/1/012036>
- Smetanin S., Komarov M. (2021) Share of Toxic Comments among Different Topics: The Case of Russian Social Networks // In: *2021 IEEE 23rd Conference on Business Informatics (CBI)*, Bolzano (Italy), September 1-3, 65–70. <https://doi.org/10.1109/CBI52690.2021.10056>
- Thelwall M., Stuart E. (2019) She's Reddit: A source of statistically significant gendered interest information? // *Information processing & management*: 56(4): 1543–58. <https://doi.org/10.1016/j.ipm.2018.10.007>

- Thorpe Huerta D., Hawkins J.B., Brownstein J.S., Hswen Y. (2021) Exploring discussions of health and risk and public sentiment in Massachusetts during COVID-19 pandemic mandate implementation: A Twitter analysis // *SSM-Population Health*: 15: 100851. <https://doi.org/10.1016/j.ssmph.2021.100851>
- Wawrzuta D., Jaworski M., Gotlib J., Panczyk M. (2021) What arguments against COVID-19 vaccines run on Facebook in Poland: Content analysis of comments // *Vaccines*: 9(5): 481. <https://doi.org/10.3390/vaccines9050481>
- Wawrzuta D., Klejdysz J., Jaworski M., Gotlib J., Panczyk M. (2022) Attitudes toward COVID-19 Vaccination on Social Media: A Cross-Platform Analysis // *Vaccines*: 10(8): 1190. <https://doi.org/10.3390/vaccines10081190>
- Xue J., Macropol K., Jia Y., Zhu T., Gelles R.J. (2019) Harnessing big data for social justice: An exploration of violence against women-related conversations on Twitter // *Human Behavior and Emerging Technologies*: 1(3): 269–79. <https://doi.org/10.1002/hbe2.160>

Другие источники информации

- Jensen E. (2016) NPR website to get rid of comments // National Public Radio, August 17. URL: <https://www.npr.org/sections/publiceditor/2016/08/17/489516952/npr-website-to-get-rid-of-comments>

Приложение

Таблица П-1. Только репродуктивные данные, 2 класса («0» — нет аргумента, «1» — есть)

	Precision	Recall	F1-score	Число наблюдений в валидационной выборке
Класс «0»	0,73	0,91	0,81	324
Класс «1»	0,79	0,50	0,61	218
Accuracy			0,75	542
Macro avg	0,76	0,70	0,71	542
Weighted avg	0,76	0,75	0,73	542

Источник: составлено авторами.

Таблица П-2. Репродуктивные и данные COVID-19 только в обучающей выборке, 2 класса («0» — нет аргумента, «1» — есть аргумент)

	Precision	Recall	F1-score	Число наблюдений в валидационной выборке
Класс «0»	0,72	0,94	0,82	346
Класс «1»	0,77	0,36	0,49	195
Accuracy			0,73	541
Macro avg	0,75	0,65	0,66	541
Weighted avg	0,74	0,73	0,70	541

Источник: составлено авторами.

Таблица П-3. Предобучение на данных COVID-19, потом на репродуктивных, 2 класса

	Precision	Recall	F1-score	Число наблюдений в валидационной выборке
Класс «0»	0,68	0,93	0,79	324
Класс «1»	0,77	0,35	0,48	218
Accuracy			0,70	542
Macro avg	0,73	0,64	0,64	542
Weighted avg	0,72	0,70	0,66	542

Источник: составлено авторами.

Таблица П-4. Только репродуктивные данные, но в урезанном виде (отобраны следующим образом: ТОЛЬКО если не менее 2 из 3 аннотаторов согласились, что есть довод), 2 класса («1» — личное по мнению не менее 2 из 3 аннотаторов, «0» — все остальное)

	Precision	Recall	F1-score	Число наблюдений в валидационной выборке
Класс «0»	0,94	0,79	0,86	202
Класс «1»	0,61	0,87	0,71	76
Accuracy			0,81	278
Macro avg	0,77	0,83	0,79	278
Weighted avg	0,85	0,81	0,82	278

Источник: составлено авторами.

Таблица П-5. Только репродуктивные данные, но в урезанном виде (отобраны следующим образом: ТОЛЬКО если не менее 2 из 3 аннотаторов согласились, что есть довод), 2 класса («1» — общественное по мнению не менее 2 из 3 аннотаторов, «0» — все остальное)

	Precision	Recall	F1-score	Число наблюдений в валидационной выборке
Класс «0»	0,62	0,80	0,70	94
Класс «1»	0,88	0,76	0,81	184
Accuracy			0,77	278
Macro avg	0,75	0,78	0,76	278
Weighted avg	0,79	0,77	0,78	278

Источник: составлено авторами.

Таблица П-6. Только репродуктивные данные, 3 класса («0» — нет аргумента, «1» — общественное, «2» — личное)

	Precision	Recall	F1-score	Число наблюдений в валидационной выборке
Класс «0»	0,72	0,85	0,78	324
Класс «1»	0,57	0,48	0,52	123
Класс «2»	0,46	0,27	0,34	95
Accuracy			0,67	542
Macro avg	0,59	0,54	0,55	542
Weighted avg	0,64	0,67	0,65	542

Источник: составлено авторами.

Информация об авторах

- Калабихина Ирина Евгеньевна — доктор экономических наук, заведующая кафедрой народонаселения, экономический факультет МГУ имени М.В. Ломоносова, Москва, 119991, Россия. E-mail: kalabikhina@econ.msu.ru
- Зубова Екатерина Андреевна — участник программы «Fox International Fellowship», Йельский университет, Нью-Хейвен, Коннектикут, 06511, США. E-mail: ekaterina.zubova@yale.edu
- Лукашевич Наталья Валентиновна — доктор технических наук, ведущий научный сотрудник, НИВЦ МГУ имени М.В. Ломоносова, Москва, 119899, Россия. E-mail: louk_nat@mail.ru
- Колотуша Антон Васильевич — кандидат экономических наук, программист 2 категории, экономический факультет МГУ имени М.В. Ломоносова, Москва, 119991, Россия. E-mail: koav13a@econ.msu.ru
- Казбекова Зарина Германовна — научный сотрудник кафедры народонаселения и лаборатории исследований цифровой экономики, экономический факультет МГУ имени М.В. Ломоносова, Москва, 119991, Россия. E-mail: kazbekova@econ.msu.ru
- Банин Евгений Петрович — кандидат технических наук, научный сотрудник, НИЦ Курчатовский институт, Москва, 123182, Россия. E-mail: evg.banin@gmail.com
- Клименко Герман Андреевич — аспирант, экономический факультет МГУ имени М.В. Ломоносова, Москва, 119991, Россия. E-mail: german89000@mail.ru