DOI: 10.14515/monitoring.2025.4.2996





И. Е. Калабихина, З. Г. Казбекова, В. С. Мошкин, М. И. Кашин, М. М. Таипов

КТО И ПОЧЕМУ ОТКАЗЫВАЕТСЯ ОТ ТАБАКОКУРЕНИЯ В РОССИИ (НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНЫХ МЕДИА И ПРИМЕНЕНИЯ НЕЙРОСЕТЕЙ)

Правильная ссылка на статью:
Калабихина И. Е., Казбекова З. Г., Мошкин В. С., Кашин М. И., Таипов М. М. Кто и почему отказывается от табакокурения в России (на основе данных социальных медиа и применения нейросетей) // Мониторинг общественного мнения: экономические и социальные перемены. 2025. № 4. С. 28—52. https://www.doi.org/10.14515/monitoring.2025.4.2996.
For citation:
Kalabikhina I. E., Kazbekova Z. G., Moshkin V. S., Kashin M. I., Taipov M. M. (2025) Who and Why Quits Smoking in Russia (Based on Social Media Data and the Use of Neural Networks). *Monitoring of Public Opinion: Economic and Social Changes.* No. 4. P. 28–52. https://www.doi.org/10.14515/monitoring.2025.4.2996. (In Russ.)

Получено: 24.03.2025. Принято к публикации: 10.07.2025.

КТО И ПОЧЕМУ ОТКАЗЫВАЕТСЯ ОТ ТАБАКОКУРЕНИЯ В РОССИИ (НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНЫХ МЕДИА И ПРИМЕНЕНИЯ НЕЙРОСЕТЕЙ)

WHO AND WHY QUITS SMOKING IN RUSSIA (BASED ON SOCIAL MEDIA DATA AND THE USE OF NEURAL NETWORKS)

КАЛАБИХИНА Ирина Евгеньевна — доктор экономических наук, зав. кафедрой народонаселения экономического факультета, Московский государственный университет им. М. В. Ломоносова, Москва, Россия
E-MAIL: ikalabikhina@yandex.ru
https://orcid.org/0000-0002-3958-6630

Irina E. KALABIKHINA[1] — Dr. Sci. (Econ.), Head of the Population Department at the Faculty of Economics
E-MAIL: ikalabikhina@yandex.ru
https://orcid.org/0000-0002-3958-6630

КАЗБЕКОВА Зарина Германовна — кандидат экономических наук, научный сотрудник кафедры народонаселения экономического факультета, МГУ имени М. В. Ломоносова, Москва, Россия
E-MAIL: kazbekova.zarina@bk.ru
https://orcid.org/0000-0002-7567-3184

Zarina G. KAZBEKOVA[1] — Cand. Sci. (Econ.), Researcher at the Population Department, Faculty of Economics
E-MAIL: kazbekova.zarina@bk.ru
https://orcid.org/0000-0002-7567-3184

МОШКИН Вадим Сергеевич — кандидат технических наук, проректор по цифровой трансформации, Ульяновский государственный технический университет, Ульяновск, Россия
E-MAIL: v.moshkin@ulstu.ru
https://orcid.org/0000-0002-9258-4909

Vadim S. MOSHKIN[2] — Cand. Sci. (Tech.), Vice-Rector for Digital Transformation
E-MAIL: v.moshkin@ulstu.ru
https://orcid.org/0000-0002-9258-4909

КАШИН Максим Игоревич — младший научный сотрудник Департамента научных исследований , Ульяновский государственный технический университет, Ульяновск, Россия
E-MAIL: m.kashin@ulstu.ru
https://orcid.org/0009-0007-4547-8306

Maksim I. KASHIN[2] — Junior Researcher, Department of Scientific Research
E-MAIL: m.kashin@ulstu.ru
https://orcid.org/0009-0007-4547-8306

ТАИПОВ Михаил Маратович — аспирант, инженер кафедры математических методов анализа экономики экономического факультета, МГУ имени М. В. Ломоносова, Москва, Россия
E-MAIL: mmtaipov@yandex.ru
https://orcid.org/0009-0003-5375-3663

Mikhail M. TAIPOV[1] — Postgraduate Student, Engineer at the Department of Mathematical Methods of the Economics Analysis, Faculty of Economics
E-MAIL: mmtaipov@yandex.ru
https://orcid.org/0009-0003-5375-3663

[1] Lomonosov Moscow State University, Moscow, Russia
[2] Ulyanovsk State Technical University, Ulyanovsk, Russia

DOI: 10.14515/monitoring.2025.4.2996

Аннотация. Цель работы — выявить специфику мотивации бросать или не бросать курить среди русскоязычных пользователей социальных медиа. Авторы исследуют систему мнений русскоязычных пользователей социальных сетей по вопросам самосохранительного поведения на основе тематического анализа контента социальных медиа с использованием больших языковых моделей (LLM). Сформированный для этих целей датасет включает более 58 тыс. комментариев на русском языке. Источник комментариев — дискуссии под видеороликами по теме курения, вручную отобранными авторами в русскоязычном YouTube-сегменте.

В ходе исследования разработан и апробирован алгоритм классификации доводов пользователей социальных медиа по вопросам мотивации табакокурения и мотивации отказа от табакокурения по заданной авторами типовой структуре; разработаны и апробированы алгоритмы классификации определения пола и возраста автора комментария на русском языке социальной сети; построены распределения причин (не)отказа от табакокурения пользователей социальных медиа, в том числе в разрезе демографических характеристик пользователей — пола и возраста. При анализе полученного массива комментариев авторы показывают, что основными доводами в пользу отказа от курения оказываются здоровье и экономия денег, причем первый встречается вдвое чаще второго; среди доводов к сохранению этой привычки выделяются опасения, связанные с лишним весом. При этом существенных гендерных и возрастных различий в доводах отказа или не отказа от курения выявлено не было.

Ключевые слова: самосохранительное поведение, табакокурение, генеративный искусственный интеллект, большие языковые модели, цифровая демография, социальные сети

Abstract. The aim of the work is to identify the specifics of motivation to quit or not to quit smoking among Russian-speaking users of social media. The authors study the system of opinions of Russian-speaking users of social networks on issues of self-preservation behavior based on thematic analysis of social media content using large language models (LLM). The dataset formed for these purposes includes more than 58 thousand comments in Russian. The comments were collected under videos on the topic of smoking, manually selected by the authors in the Russian-language YouTube segment.

The study presents and tests an algorithm for classifying arguments of social media users on issues of motivation for smoking and motivation for quitting smoking, develops and validates algorithms for classifying the gender and age of the author of a comment, and construts distributions of reasons for (not) quitting smoking in general and by demographic characteristics of users (gender and age). The analysis of the compiled dataset showed that the main arguments in favor of quitting smoking are health and saving money, with the former occurring twice as often as the latter; among the arguments for maintaining this habit, concerns related to excess weight stand out. At the same time, no significant gender and age differences in the arguments for quitting or not quitting smoking were revealed.

Keywords: self-preservation behavior, smoking, generative artificial intelligence, large language models, digital demography, social networks

Благодарность. Исследование выполнено в рамках НИР «Воспроизводство населения в социально-экономическом развитии» 122041800047-9 и внутреннего гранта экономического факультета МГУ им. М. В. Ломоносова на тему: «Демографические детерминанты оценки качества медицинских услуг и отказа от табакокурения: анализ мнений россиян на основе применения нейросетей и генеративного искусственного интеллекта». Авторы выражают благодарность коллегам, участвовавшим в разметке массивов комментариев пользователей социальных медиа: Антону Колотуше и Софии Журавлевой.

Acknowledgments. The study was conducted within the framework of the research project "Population Reproduction in Socio-Economic Development" 122041800047-9 and an internal grant from the Faculty of Economics of Lomonosov Moscow State University "Demographic Determinants of Assessing the Quality of Medical Services and Smoking Cessation: Analysis of Russians' Opinions Based on the Use of Neural Networks and Generative Artificial Intelligence". The authors thank their colleagues Anton Kolotusha and Sofia Zhuravleva who participated in annotating the arrays of social media user comments.

Введение

Один из ключевых вопросов в демографических исследованиях — состояние здоровья населения, а также влияющие на него факторы и возможности улучшения ситуации в этой области. В настоящей работе мы обращаемся к данным, касающимся вопросов здоровья населения, — это мнения людей о том, почему надо (не) бросать курить. С чем связан такой выбор? Здоровье человека зависит от ряда факторов. Как правило, выделяют триаду детерминант здоровья: качество системы здравоохранения, поведенческий фактор, качество среды жизнедеятельности (материальное благополучие, экологический фон и пр.). Первые два фактора находятся непосредственно в поле зрения политики в области повышения уровня здоровья населения. Последний фактор опосредованно влияет на здоровье, он более комплексный и трудно управляемый. Мы изучаем поведенческий фактор — самосохранительное поведение, которое может быть направлено как на сбережение здоровья, так и на его разрушение.

В рамках данного исследования мы выполняем задачу по разработке и апробации методологии мониторинга двух типов доводов в области самосохранительного поведения: доводы бросить курить и доводы не бросать курить.

Изучение самосохранительного поведения в области вредных привычек актуально в период проведения семейной и демографической политики (2007—2025 гг.), антитабачной политики (особенно активна с 2013 г.) и на фоне отстающего от среднего по миру темпа снижения распространенности курения в России (см. рис. 1). Выявление гендерных особенностей в мотивации бросать или не бросать курить среди русскоязычных пользователей социальных медиа также обладает высокой актуальностью — в связи с разнонаправленной динамикой распространенности курения среди женщин и мужчин в России (см. рис. 2).

Разработанные нами алгоритмы классификации доводов (не) бросать курить позволят при наличии мониторинга в режиме реального времени получать информацию о том, какой из основных факторов вынуждает россиян бросить курить в большей степени (например, вред для здоровья или дороговизна сигарет)

и насколько в российском обществе распространены те или иные мифы о вреде прекращения курения, причем ответы на эти вопросы будут в разрезе отдельных демографических групп. Такие данные могут использоваться для аргументации мер демографической политики в перспективе: в зависимости от полученных результатов меры политики по борьбе с курением могут быть настроены более эффективно, а значит, быстрее и эффективнее приведут к конечной цели — снижению распространенности курения в стране.

Рис. 1. Распространенность курения в России и мире, в %[1]

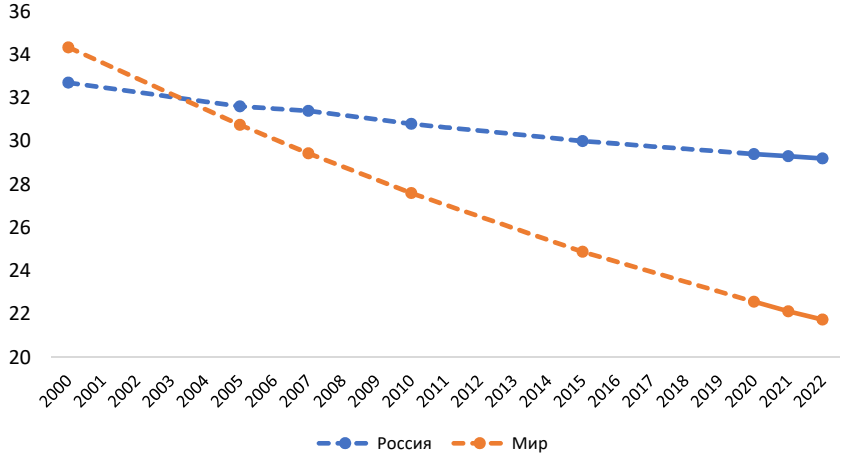
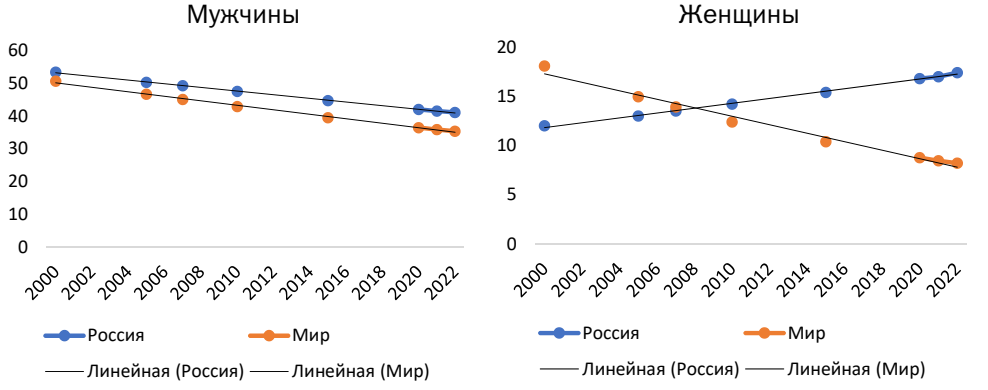


Рис. 2. Распространенность курения среди мужчин и женщин в России и мире, в %[2]



[1] Источник: построено авторами по данным Всемирного банка.

[2] Источник: построено авторами по данным Всемирного банка.

Курение табака имеет крайне негативные последствия для общественного здоровья. Только в 2019 г. в мире преждевременно ушли из жизни 7,69 млн курильщиков табака. Более того, курение табака уносит жизни не только самих курильщиков: в том же году в мире скончалось 1,3 млн человек, не куривших табак, но вдыхавших табачный дым [Ritchie, Roser, 2023]. Поэтому правительства стран заинтересованы в сокращении потребления табачных продуктов. Для достижения данной цели необходима продуманная политика, которая невозможна без верного понимания отношения населения к курению табака.

Изучение отношения к курению и доводов (не)отказа от курения с использованием данных социальных сетей и методов машинного обучения

Методы машинного обучения регулярно используются для изучения отношения к курению и поведения курильщиков. На основе анализа записей в соцсетях или опросов общественного мнения методами машинного обучения были определены ключевые факторы, влияющие на вероятность отказа от курения: наличие курящих домочадцев, доступность сигарет, уровень образования, недавнее употребление алкоголя, вступление в новые отношения, переезд и выход на новую работу, наличие запрета на курение [Culotta, 2010; Coughlin et al., 2020; Bickel et al., 2023]. В нашей работе мы исследуем аргументы, приводимые в пользу или против отказа от табакокурения, а курение электронных сигарет (вейпинг) и кальяна рассматриваем только как распространенные способы отказа от курения табака[3]. Однако представляет интерес анализ работ, в которых методы машинного обучения использовались для выявления отношения людей к вейпингу или курению кальяна. Ш. Висвесваран с коллегами [Visweswaran et al., 2020] отобрали твиты о курении электронных сигарет, при помощи моделей глубокого обучения классифицировали некоммерческие твиты по позитивному или негативному отношению к электронным сигаретам. Результаты работы показали, что 62,39 % некоммерческих публикаций содержали положительное отношение к курению электронных сигарет. К.-Х. Чу с коллегами [Chu et al., 2019] использовали модели машинного обучения для анализа твитов о курении кальяна. Было обнаружено, что большинство твитов содержали положительное отношение к употреблению кальяна.

Поскольку наша цель — классификация аргументов, приводимых в поддержку или против отказа от табакокурения, наибольший интерес для нас представляют статьи, в которых методы машинного обучения используются для нахождения и классификации аргументов, приводимых в различных дискуссиях. Р. Кавулуру и А. Саббир [Kavuluru, Sabbir, 2016] на основе метода логистической регрессии (LR) создали модель, способную решать задачу классификации сторонников вейпинга по их твитам. Однако среди методов машинного обучения, которые могут быть использованы для классификации доводов отказа от курения, особый интерес для нас имеют большие языковые модели (LLM, Large Language Models), поскольку их структура хорошо подходит для анализа аргументов в дискуссиях.

[3] Рынок альтернативных способов курения активно развивается в мире. Согласно отчету NielsenIQ, в России в 2022 г. среди продуктов для курения наиболее динамично рос рынок одноразовых электронных сигарет. См. Тренды индустрии: рынок табачных изделий // Nielsenl Q. 2022. 25 июля.. URL: https://nielseniq.com/global/ru/insights/analysis/2022/trendy-industrii-rynok-tabachnyh-izdeliy (дата обращения: 01.09.2025).

LLM — это нейронные лингвистические сети, имеющие трансформенную архитектуру, обученные на больших корпусах данных и решающие задачи анализа текстовой информации: классификации, генерации, резюмирования и пр. Главное преимущество больших языковых моделей — возможность семантической оценки текста, то есть его «понимание» посредством учета контекста и семантической близости минимальных структурных единиц текста (токенов). М. Гуида с коллегами использовали LLM для анализа аргументов, содержащихся в текстах с дискуссионных платформ. LLM были применены для решения следующих трех задач: установления того, используется ли заданный аргумент в комментарии; извлечения куска текста, содержащего этот аргумент; определения того, приводится ли аргумент в поддержку позиции автора или автор критикует этот аргумент. Получен вывод, что LLM хорошо справляются с первой и третьей задачами, но мало пригодны для извлечения кусков текстов, содержащих аргументы [Guida et al., 2025]. LLM в некоторых работах используются для определения отношения к табачным продуктам [Kim, Kim, 2025], но способность LLM проверять наличие заданных аргументов в тексте указывает на то, что они также могут быть использованы для классификации аргументов в пользу отказа от табакокурения.

Классификация комментариев в социальных сетях об отказе от курения по демографическим характеристикам автора

Отношение к курению может заметно различаться среди разных демографических групп. Женщины чаще всего не способны бросить курить из-за таких факторов, как стресс и страстная тяга к курению, а мужчины — из-за повсеместной доступности сигарет и давления окружения [Dieleman, van Peet, Vos, 2021]. Поэтому, чтобы разработать научно обоснованную доказательную политику, важно иметь информацию о поле (и возрасте) курильщиков. К сожалению, не всегда возможно получить доступ к данным характеристикам автора отзыва или комментария в социальных сетях[4]. По этой причине в отдельных исследованиях ученые использовали методы машинного обучения для того, чтобы на основе текстов на различных языках определить пол или возраст их авторов.

Н. Ченг, Р. Чандрамули и К. П. Суббалакшми для определения пола применяли классификационные модели машинного обучения, использующие в качестве аргументов различные текстовые характеристики [Cheng, Chandramouli, Subbalakshmi, 2011]. Авторы обнаружили значительные стилистические различия между текстами, написанными мужчинами и женщинами, которые позволяют распознать пол их авторов. Х. Химди и К. Шаалан для определения пола авторов арабских текстов построили классификационные модели, основанные на различных методах машинного обучения [Himdi, Shaalan, 2024]. Результаты анализа показали, что наибольшей точностью обладает модель, которая использует метод BERT и учитывает текстовые особенности, свойственные лицам определенного пола. Подобная модель обладает точностью 91%. В. Юнкин, М. Литвак и И. Рабаев были исследовали литературные тексты на английском языке при помощи различных методов машинного обучения [Younkin, Litvak, Rabaev, 2024]. Авторы обнаружи-

[4] Большинство баз и соцсетей не содержат подобных данных, а «ВКонтакте» и X содержат их с определенными оговорками.

ли, что при помощи таких методов, как генеративный предобученный трансформер 2 (GPT2), и метода XLNET, соединяющего сильные стороны GPT2 и нейросетевой языковой модели на архитектуре трансформера (BERT), можно добиться точности прогнозов выше 90 %. Классификационные модели машинного обучения также применяются для определения пола авторов текстов на русском языке [Sboev et al., 2016; Sboev et al., 2018; Сбоев и др., 2023].

Методы машинного обучения могут помочь определить не только пол, но и возраст авторов текстов. А. Кляйн, А. Магги, Г. Гонзалес-Хернандес [Klein, Magge, Gonzalez-Hernandez, 2022] на основе метода NLP создали модель, способную определить точный возраст пользователей соцсети X на основе их собственных заявлений. Модель сумела предсказать возраст 54 % пользователей. F1-значений данных предсказаний составляет 0,855. А. Романов с коллегами [Romanov et al., 2020] при помощи моделей, основанных на глубоких нейронных сетях, провели анализ записей на русском языке в социальной сети «ВКонтакте». Наибольшей точностью обладает модель на основе архитектуры FastText, способная правильно определить возраст авторов 82 % постов.

К. О'Коннор и соавторы [O'Connor et al., 2024] на основе анализа статей, исследующих определение либо пола авторов записей в соцсети X, либо их возраста, либо пола и возраста одновременно, сделали следующие выводы. Определение пола — более популярная тема исследования, чем определение возраста; точность классификации пола в работах варьировалась от 51 % до 97 %, а возраста — от 43 % до 86 %.

В данной работе мы классифицируем доводы в пользу и против отказа от традиционного курения табака, представленные в дискуссиях в социальных сетях. При этом мы собираемся определить, какие из этих доводов чаще или реже встречаются среди представителей различных демографических групп. К примеру, отдельные аргументы в пользу отказа от курения могут часто приводиться женщинами, но намного реже — мужчинами. Так, только женщинам свойственно аргументировать отказ от курения боязнью набрать вес [Кузнецова, 2019]. Поскольку в ходе обзора литературы нами обнаружено, что методы машинного обучения могут с высокой точностью классифицировать типы доводов, приводимых в текстах, и определять пол и возраст их авторов, то для решения данных задач мы используем модели LLM и LSTM. Полученные выводы позволят определить, какие аргументы против курения табака находят больший отклик среди различных демографических групп, что поможет увеличить эффективность антитабачной политики.

Данные

Мы исследуем структуру доводов в пользу или против отказа от табакокурения с применением больших языковых моделей на материалах платформы YouTube — текстов комментариев по теме на русском языке. Имея опыт анализа демографического поведения россиян на основе данных во «ВКонтакте» [Kalabikhina et al., 2023; Калабихина и др., 2023], а также понимая все плюсы данных этой социальной сети для решения поставленной задачи (наличие информации о демографических характеристиках пишущего комментарий), мы начали поиск материалов именно в этой социальной сети. Поиск осуществлялся разными способами:

отбор релевантных групп (сообществ) и сбор комментариев в них, поиск релевантных комментариев (без привязки к группам). Однако ни один из способов не дал результатов: необходимого объема комментариев (от 10 тыс.) во «ВКонтакте» получить не удалось. Предварительный ручной поиск комментариев по теме в мессенджере Telegram, в социальной сети «Одноклассники», на видеохостингах Rutube и (VK Видео) также оказался безрезультатным — доступных релевантных комментариев было недостаточно для решения поставленной задачи. В итоге основным источником данных стал видеохостинг YouTube. При отборе видео на YouTube мы ориентировались на следующие критерии: 1) релевантность названия видео (видео должно быть по теме курения); 2) язык (русский язык); 3) популярность (число просмотров) — в первую очередь важно было найти наиболее популярные видео, так как они содержат наибольшее число комментариев и вовлекают большее число людей в обсуждение проблемы. Релевантные видео отбирались непосредственно на самой платформе YouTube через функцию поиска. В качестве поискового запроса использовались следующие ключевые слова: «курение», «электронные», «как я бросил курить». Итоговая база включает 204 видео, это наиболее популярные русскоязычные видео по теме курения по состоянию на май 2024 г. Датасет со скачанными нами комментариями под видео из базы размещен в открытом доступе [Kalabikhina, Kazbekova, Moshkin, 2025]. Он включает 165 тыс. комментариев.

В связи с техническими ограничениями программной библиотеки, обеспечивающей доступ к API применяемых в ходе исследования больших языковых моделей (LLM), основную работу по классификации мы выполнили по сокращенному датасету из 58 тыс. комментариев из сформированного нами датасета. Полученное множество комментариев («малый» датасет) было сформировано путем упорядочения исходного множества по дате в порядке убывания и выбора наиболее близких к текущей дате анализа комментариев. Мы классифицировали доводы и демографические характеристики авторов (пол и возраст авторов) на основе «малого» датасета. Список отобранных видео, полный датасет (с эмоциональным фоном комментариев), «малый» датасет с классифицированными по наличию и типу аргументов комментариями, а также датасет с классифицированными по полу и возрасту комментариями, имеющими аргумент (не)отказа от табакокурения (5,5 тыс. комментариев), представлены в открытом доступе [Kalabikhina, Kazbekova, Moshkin, 2025].

Методология

Мы решаем задачу по разработке и апробации методологии мониторинга двух типов доводов в области самосохранительного поведения. Структура доводов обсуждалась в ранней работе [Калабихина, Казбекова, Зубова, 2024]. Первый тип — доводы бросить курить. В этой части мы определяем, почему, по мнению россиян, следует бросить курить. Второй тип доводов — доводы не бросать курить. В данном случае мы определяем, почему, по мнению россиян, не следует бросать курить. Разработанная нами типовая структура доводов выглядит следующим образом:

1. Доводы бросить курить:

1а) здоровье (пример: «курить — вредно для здоровья»);

1б) деньги (пример: «курить — дорого»);
1в) иное.
2. Доводы не бросать курить (или антидоводы бросить курить):
2а) лишний вес (пример: «если бросить курить, то появится лишний вес»);
2б) иное.

Среди доводов бросить курить в рамках данного проекта мы выделяем две наиболее распространенные, по нашему мнению, категории: связанные с вредом для здоровья и «вредом» для кошелька курильщика. При этом все остальные доводы мы также учитываем, собирая их в отдельную категорию «иное», что позволяет нам оценивать не только взаимное соотношение выбранных двух категорий, но и их вес в общем массиве доводов.

В качестве барьеров на пути к отказу от курения мы выделяем лишний вес. Гипотеза о страхе набрать лишний вес как доводе не бросать курить основана на результатах эконометрического исследования кафедры народонаселения Экономического факультета МГУ М. В. Ломоносова о факторах приверженности табакокурению — влиянии табака на вес человека [Кузнецова, 2019].

В настоящей работе мы не только определяем вес каждого из факторов в общей структуре доводов бросить курить / барьеров, препятствующих отказу от курения, но и выявляем гендерные и возрастные особенности в полученных результатах.

В своем предыдущем исследовании [Калабихина, Казбекова, Зубова, 2024] для решения задачи классификации доводов пользователей социальных медиа по вопросам в области самосохранительного поведения (мотивация курения либо отказа от курения) мы применяли метод обработки естественного языка на основе нейромодели Conversational RuBER T. Здесь же мы реализовали классификацию с использованием возможностей генеративных моделей искусственного интеллекта — больших языковых моделей (LLM). И, что важно, мы классифицируем все типы доводов и наличие доводов в тексте одновременно (ранее мы выполняли последовательную классификацию в наборе экспериментов по каждому типу доводов).

Мы выполняем классификацию комментариев из выборки по шести типам:

1 класс — комментарии, не содержащие довод бросить курить или довод не бросать курить;

2 класс — комментарии, содержащие довод бросить курить по причине нанесения вреда здоровью курильщика;

3 класс — комментарии, содержащие довод бросить курить по причине дороговизны сигарет / экономии денежных средств;

4 класс — комментарии, содержащие довод бросить курить по иной причине (помимо заботы о своем здоровье и экономии денег);

5 класс — комментарии, содержащие довод не бросать курить из-за страха набрать лишний вес;

6 класс — комментарии, содержащие довод не бросать курить по иным причинам (помимо лишнего веса).

Примеры комментариев по каждому из рассматриваемых классов представлены в таблице 1.

Таблица 1. ***Примеры комментариев из датасета***

Класс	Пример
Класс 1 — Нет довода	«Курить нужно бросать резко». «То чувство когда смотришь видео и куришь😂».
Класс 2 — Есть довод (собственное здоровье)	«Курю 3 года недавно начались проблемы с желудком (возможно начальные этапы язвы) Особенно болит живот после скуренной сигареты, решил полностью бросить ради своего же здоровья пожелайте мне удачи чтобы я смог». «Как–то ночью, проснулась — голова болит. Давление померила: 212 на 126. Вся тяга курить пропала».
Класс 3 — Есть довод (деньги)	«Я бросил курить. Я хочу купить машину, поэтому жестко экономлю на многом. Я посчитал что в год на сигареты уходит 27—30 т. р. Может для кого то это не деньги? Но это только затраты на сигареты, а представьте еще сколько на разные вкусняшки, газировки, пиво денег уходит. За год можно на подержанную машину сэкономить». «Тоже думаю бросить курить. Слишком дорого сигареты стоят».
Класс 4 — Есть довод (иное)	«Бросил курить чтобы не подавать плохой пример ребенку! Держусь 10 дней! Примерно 5 лет назад перешел на электронку в попытках бросить, так вот электронку бросить еще сложнее чем сигареты, так как в примерном перещете на сигараты я курил три пачки в день! Так что первые три дня чуть помер), сейчас легче но еще тяжко, но ничего прорвемся) Всем кто пытается бросить курить, не ищите альтернативу, бросайте сразу что бы не вышло как у меня с пачки на три! Всем удачи)». «в середине января узнал что стану отцом, 28го января последний раз затянулся)))) курил со школы, более 25 лет, иногда ломает, но терпимо, я на свое чадо этим дерьмом дышать не буду…»
Класс 5 — Есть антидовод (лишний вес)	«Забыл сказать про набор лишнего веса. Это реально. Крение подавляет апетит. Когда бросаешь начинается жор. Очень трудно с ним бороться». «расскажите как не начать набирать вес бросив курить, потому что я знаю людей, которые поправились на 25 кг. бросив курить, а это тоже опасно для здоровья».
Класс 6 — Есть антидовод (иное)	«куришь умрешь, не куришь умрешь 🤣». «Мой бывший знакомый не когда не курил и не пил крепче кофе. И уже на том свете 😉 А ему было только 51 год 😕 А кто-то курит и бухает с 14 лет и живут до 90 лет И живей всех живых. Так что это все бла-бла». «Я бросал курить и некурил 10 месяцев некакий изменений в состоянии здоровья и внешних улучшениях я не заметил». «Что происходит, когда бросаешь курить? Лично у меня — проблемы с желудком начались (Ибо питаться стал больше). Снова начал курить — проблема ушла». «Бросил курить, заработал алергию». «курил 45 лет, бросил легко, через год получил инфаркт». «Бросив курить, несколько моих знакомых в возрасте 40+ получили бонусом диабет, все агитаторы за бросание о таких побочках умалчивают(». «а Черчилль прожил 90 лет всю жизнь курил по 30 сигар в день». «Когда бросаешь курить — начинаешь пить». «Что мне делать??? Я боюсь стать занудным и скучным из–за того, что брошу курить, и что не будет креативности и расслабления больше». «Кто не курит и не пьет тот здоровеньким помрет ☝🤢😆😆😆».

Примечание: комментарии даны в оригинальном виде, с сохранением орфографии и пунктуацим авторов.

На предварительном этапе мы провели ряд экспериментов с нейромоделями и генеративным искусственным интеллектом (LLM) для определения лучшего способа классификации данных нашего типа. Общий вывод экспериментов с различными нейросетями, с генеративным искусственным интеллектом, с комбини-

рованными подходами: наилучшей моделью классификации наличия и типов доводов (не)отказа от курения табака является использование LLM.

В связи с техническими ограничениями анализа мы выполнили классификацию 58 тыс. комментариев из сформированного нами датасета, включающего 165 тыс. комментариев.

Комментарии классифицировались с использованием модели LLM gemma2—9b—it. Для выполнения классификации были сформулированы и протестированы несколько вариантов промтов. На основе сравнения их точности был выбран наилучший вариант (общий процент совпадений составил 70 %):

'Ты будешь классифицировать комментарии по теме «Отказ от курения». Ответ нужно представить в виде цифры без объяснений. Используй следующие правила:' +

'\n1 — Если в комментарии нет ни довода, ни антидовода про отказ от курения.' +

'\n2 — Если в комментарии содержится довод бросить курить, чтобы улучшить здоровье. Примеры: «Язык очищался 6 недель, и вся гадость вышла из легких», «Здоровый дух в здоровом теле.»' +

'\n3 — Если в комментарии содержится довод бросить курить, чтобы сэкономить деньги. Примеры: «Подсчитал, сколько трачу на сигареты, и решил бросить», «В месяц уходит около 13 тыс. рублей, пора завязывать.»' +

'\n4 — Если в комментарии содержится довод бросить курить по другой причине. Примеры: «Не курю пятый месяц, и чувствую себя отлично», «Психолог подсказал, как бросить, и теперь я свободен.»' +

'\n5 — Если в комментарии содержится антидовод не бросать курить, чтобы не потолстеть. Примеры: «Проблема не в том, чтобы бросить курить, а в том, что потом начинаешь толстеть.»' +

'\n6 — Если в комментарии содержится антидовод не бросать курить по другой причине. Примеры: «Я бросил курить, но не смог спать и работать нормально», «Курение — это зависимость, и ее трудно преодолеть.»'

Результаты оценки доводов в пользу или против отказа от табакокурения с использованием большой языковой модели

Классификация доводов за или против отказа от табакокурения

Результаты классификации доводов представлены в таблице 2.

Мы получили следующие результаты. Во-первых, искомые доводы содержатся в 16 % комментариев датасета. Это примерно 10 тыс. комментариев из общей выборки в 58 тыс. комментариев. Данный результат является нормальным. Во-вторых, при формировании выборки мы собираем *все* комментарии под соответствующими видео по теме курения, не проводя их предварительную обработку. В-третьих, наличие довода определяется с очень высокой точностью. В-четвер-

тых, отдельные доводы классифицируются с разной степенью точности. Распределение доводов бросить курить представлено на рисунке 3.

Таблица 2. ***Результаты LLM классификации комментариев по доводам и анти-доводам отказа от курения***

Класс	Промт	Число комментариев (шт.)	Число комментариев (%)
Класс — Ошибка классификации	Во время классификаций произошла ошибка	1 328	2,29
Класс 1 — Нет довода	Если в комментарии нет ни довода, ни антидовода про отказ от курения	47 297	81,55
Класс 2 — Есть довод (собственное здоровье)	Если в комментарии содержится довод бросить курить, чтобы улучшить здоровье. Примеры: «Язык очищался 6 недель, и вся гадость вышла из легких», «Здоровый дух в здоровом теле»	1 317	2,27
Класс 3 — Есть довод (деньги)	Если в комментарии содержится довод бросить курить, чтобы сэкономить деньги. Примеры: «Подсчитал, сколько трачу на сигареты, и решил бросить», «В месяц уходит около 13 тыс. рублей, пора завязывать»	591	1,02
Класс 4 — Есть довод (иное)	Если в комментарии содержится довод бросить курить по другой причине. Примеры: «Не курю пятый месяц, и чувствую себя отлично», «Психолог подсказал, как бросить, и теперь я свободен»	4 297	7,41
Класс 5 — Есть антидовод (лишний вес)	Если в комментарии содержится антидовод не бросать курить, чтобы не потолстеть. Примеры: «Проблема не в том, чтобы бросить курить, а в том, что потом начинаешь толстеть»	267	0,46
Класс 6 — Есть антидовод (иное)	Если в комментарии содержится антидовод не бросать курить по другой причине. Примеры: «Я бросил курить, но не смог спать и работать нормально», «Курение — это зависимость, и ее трудно преодолеть»	2 901	5,00

Чаще встречается довод, связанный с заботой о здоровье. Аномально высокий результат по классу 4 — иные доводы бросить курить. Он противоречит как интуиции, так и выводам, полученным нами на предыдущем этапе исследования, когда мы использовали нейросеть BERT и реализовывали подход, в рамках которого каждый довод классифицировался в рамках отдельного эксперимента. Мы попросили большую языковую модель дополнительно сформировать столбец с объяснением своего выбора. Это позволяет узнать, почему модель отнесла комментарий к тому или иному классу. Изучив объяснения выбора 4 класса моделью, мы сделали вывод, что она неверно понимает задачу и делает систематические ошибки в случае, когда нет четкого определения класса (классы типа «иное»). В связи с этим мы решили выполнить ручную разметку 4 297 комментариев, которые мо-

дель отнесла к четвертому классу (иные доводы бросить курить). В итоге из 4 297 комментариев к нему было отнесено лишь 366 (менее 9 %), что свидетельствует о низкой точности данной классификации в случае класса «иное» (см. рис. 4).

Рис. 3. Результаты автоматической классификации комментариев: распределение доводов бросить курить (с помощью ГИИ), шт.

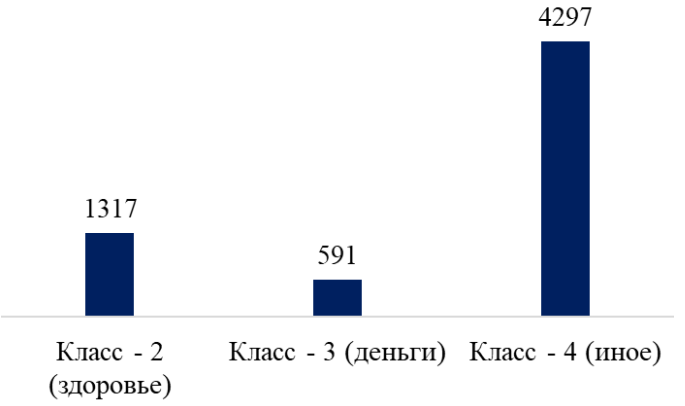
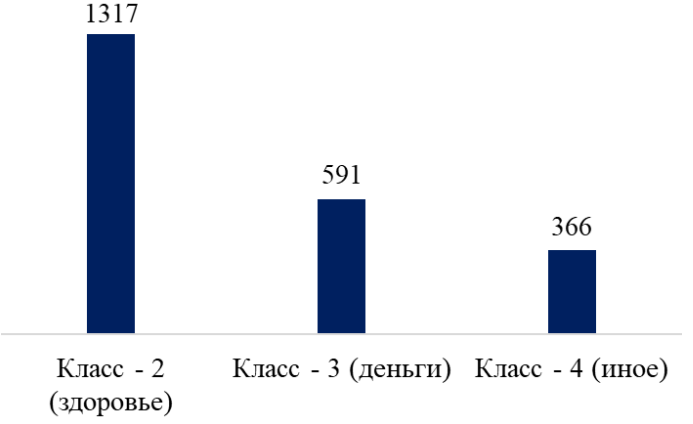


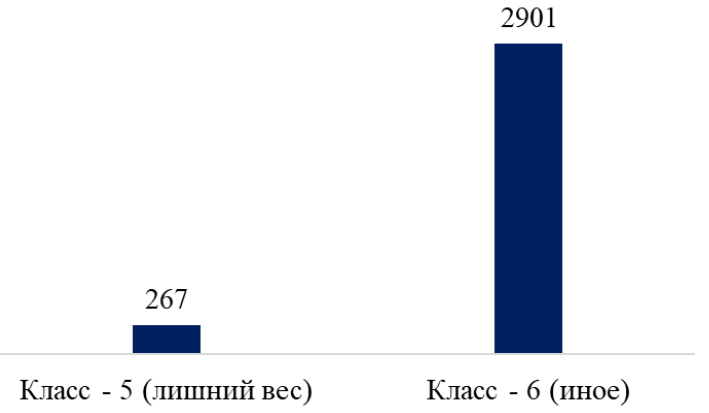
Рис. 4. Результаты классификации (класс 2 и класс 3) и ручной разметки (класс 4) комментариев: распределение доводов бросить курить, шт.



Таким образом, наиболее распространенные причины, побуждающие пользователей YouTube бросить курить, — это здоровье и деньги. Причем здоровье встречается чаще, чем деньги, примерно в два раза.

Распределение доводов не бросать курить представлено на рисунке 5. По оценке, лишний вес является одним из барьеров на пути отказа от курения: его вес в общей структуре доводов составляет 8 %.

Рис. 5. Результаты автоматической классификации (класс 5 и класс 6): распределение доводов не бросать курить (барьеров на пути к отказу от курения), шт.



В данном случае структура классификации была проще. По сути, мы вновь выделяли один класс (лишний вес) из массива. Результат классификации здесь сопоставим с нашими предыдущими подходами.

Важным ограничением используемого алгоритма является то, что он не учитывает возможные пересечения — ситуации, когда в одном комментарии содержится несколько доводов (по нашим оценкам, доля таких комментариев составляет около 9 %).

Определение пола автора комментария

В связи с тем, что на YouTube не представлена информация о социально-демографических характеристиках пользователей, таких как пол, возраст, семейное положение и т. д., мы поставили задачу разработать методологию автоматического определения пола и возраста автора комментария на основе возможностей генеративного искусственного интеллекта. Это особенно важно для российской аудитории, поскольку, как мы выяснили эконометрическими методами, женщины — в отличие от мужчин — склонны откладывать отказ от курения табака ввиду страха набрать вес [Кузнецова, 2019].

Задача автоматического определения пола автора комментария видится нам выполнимой главным образом в связи с тем, что в русском языке есть лингвистические маркеры, указывающие на пол. Это, в частности, окончания глаголов прошедшего времени в первом лице (*«Я **бросила** курить, потому что это вредило моему здоровью»*), окончания прилагательных в первом лице (*«Я еще слишком **молодая**, чтобы отравлять себе здоровье сигаретами»*) и существительные, которые пишутся различно для разных полов (*«Я **курильщица** со стажем»*). В нашей выборке чаще всего встречается первый признак.

Следует отметить наличие возможности намеренного или ненамеренного (в результате ошибки/опечатки) искажения информации о своем поле в тексте. Однако сложно представить причину, по которой автор комментария под видео

по теме курения на YouTube захотел/захотела бы притвориться человеком другого пола. Мы ожидаем, что процент таких искажений — случайных и намеренных — в выборке невысок, и поэтому данная проблема не является серьезным ограничением.

Конечно, не все комментарии содержат явные маркеры пола. Например, пол авторов следующих комментариев однозначно установить невозможно — такой комментарий может написать как мужчина, так и женщина:

— *«Не хочу я бросать курить — это снижает стресс для меня».*
— *«Я курю, чтобы не набрать вес».*

Однако часть комментариев все же содержат перечисленные маркеры. Мы вручную разметили 6 тыс. комментариев из общей выборки по следующим признакам: 1) автор комментария — мужчина; 2) автор комментария — женщина; 3) пол автора комментария установить невозможно. Мы получили следующие результаты: 1) в 68% случаев пол установить невозможно (отсутствуют явные признаки); 2) 28% — комментарии мужчин; 3) 4% — комментарии женщин. Таким образом, примерно 32% комментариев оказались с признаками, по которым можно установить пол автора текста. Если предположить, что в датасете, включающем 165 тыс. комментариев, пропорция сохранится, то мы получим около 50 тыс. комментариев, в которых можно точно установить пол автора по лингвистическим признакам текста.

Можно заключить, что теоретически искусственный интеллект может принимать обоснованные решения при определении пола автора комментария, опираясь на описанные лингвистические особенности. В связи с этим мы поставили задачу классифицировать комментарии с доводами (бросить курить и не бросать курить) по полу по трем категориям: 1) мужчина; 2) женщина; 3) пол установить невозможно. Мы решили добавить третью категорию, чтобы повысить точность результатов и не «заставлять» модель делать выбор во всех случаях. По нашему мнению, тогда модель в первую очередь классифицирует комментарии с явными лингвистическими «подсказками», о которых мы писали ранее. В таблице 4 представлены сформулированные нами промты для решения задачи классификации комментариев по полу и точность каждого из них.

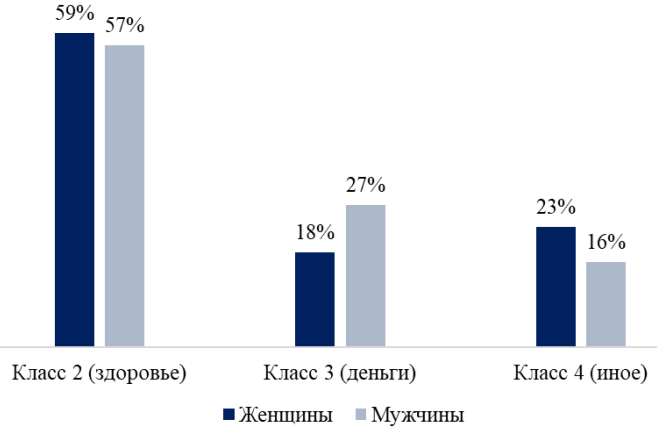
Для проверки точности промтов мы использовали уже упомянутые результаты ручной разметки 6 тыс. комментариев из общей выборки по следующим признакам: 1) автор комментария — мужчина; 2) автор комментария — женщина; 3) пол автора комментария установить невозможно.

В связи с тем, что второй промт лучше определяет пол автора комментария, мы сделали выбор в его пользу. Результаты классификации 5442 комментариев по полу авторов и доводам с использованием второго промта представлены на рисунках 6 и 7.

Таблица 4. ***Промты для автоматического определения пола и точность классификации***

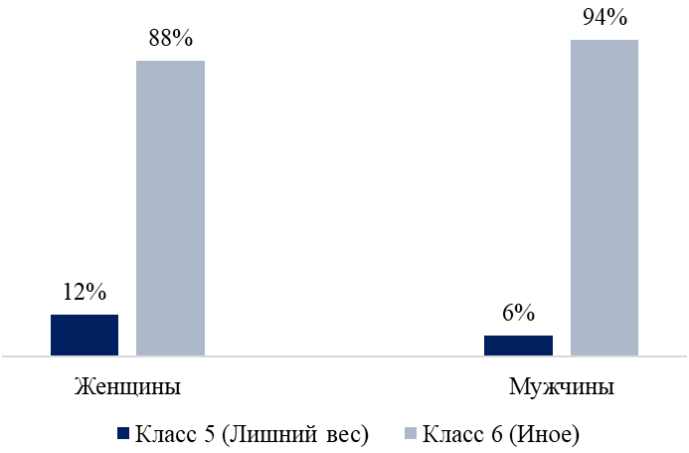
Промт	Доля верно предсказанных комментариев класса «Мужчины», %	Доля верно предсказанных комментариев класса «Женщины», %	Доля верно предсказанных комментариев класса «Невозможно определить пол», %
1) Определи пол авторов комментариев. Представь результаты в виде таблицы из 2 столбцов: первый — объяснение выбора, второй — пол автора (мужской или женский или невозможно определить)	39	52	70
2) Определи пол авторов комментариев. Представь результаты в виде таблицы из 2 столбцов: первый — объяснение выбора, второй — пол автора (мужской или женский или невозможно определить). В первую очередь обращай внимание на окончания глаголов, указывающие на принадлежность к мужскому или женскому полу	84	84	43
3) Определи пол авторов комментариев. Представь результаты в виде таблицы из 2 столбцов: первый — объяснение выбора, второй — пол автора (мужской или женский или невозможно определить). В первую очередь обращай внимание на окончания глаголов, указывающие на принадлежность к мужскому или женскому полу. Пример: «Я сделала это» — автор этого комментария женщина, это видно по форме глагола. «Я сделал это» — автор этого комментария — мужчина	83	84	60

Рис. 6. Результаты классификации комментариев по типу довода бросить курить и по полу, %



Оцененная структура причин, побуждающих бросить курить мужчин и женщин, достаточно схожа, хотя есть и отличия (ср. рис. 4). В структуре обоих полов преобладающее место занимает здоровье (59 % и 57 % для женщин и мужчин соответственно). Деньги — более важный фактор для мужчин: его вес составляет 27 % для мужчин и 18 % — для женщин. Вес всех остальных доводов бросить курить составляет, по оценке, 16 % для мужчин и 23 % — для женщин. Напомним, что в категории иных доводов бросить курить в нашем датасете встречаются следующие: 1) ответственность перед детьми, в том числе за их здоровье; 2) больше свободного времени у тех, кто не курит; 3) вред, который курильщик наносит другим людям / окружающей среде; 4) низкое качество современных сигарет; 5) религиозные настроения, несовместимые с табакокурением; 6) запах сигарет.

Рис. 7. Результаты классификации комментариев по типу довода не бросать курить и по полу, %



Оцененная нами структура причин, препятствующих отказу от курения, у мужчин и женщин также достаточно схожа (см. рис. 7): для представителей обоих полов лишний вес не является наиболее важным фактором, однако для женщин он имеет больший вес, чем для мужчин (12 и 6 % соответственно).

В таблице 1 Приложения представлены построенные структуры с распределением абсолютного числа комментариев каждого класса.

Определение возраста автора комментария

По сравнению с полом лингвистических признаков, по которым можно достоверно определить возраст или хотя бы укрупненную возрастную группу человека, написавшего текст, намного меньше. Хотя маркеры все-таки возможны. Например, авторов следующих комментариев с определенной долей уверенности можно отнести к категории молодежи:

— *Многие **мои одноклассники** курят, поэтому и я тоже стал курить.*
— ***Нам в школе** запрещено курить, **мама** мне тоже **запрещает** — поэтому и бросил.*

В некоторых случаях комментатор может указать свой возраст в тексте, однако таких случаев практически нет.

Отмеченные сложности побудили нас сконцентрироваться на одной возрастной группе — молодежь/подростки. Мы выделили три класса: до 18 лет, 19—34 года, 35+ лет.

Для проверки точности моделей мы использовали датасет с комментариями из социальной сети «ВКонтакте» по вопросам репродуктивного поведения, сформированный нами на прошлых этапах настоящего исследования. Он включает 5 тыс. комментариев. Диапазон возрастов: 15—49 лет.

Он имеет следующие характеристики:

1 класс (до 18 лет) — 519 строк;

2 класс (от 19 до 34 лет) — 3950 строк;

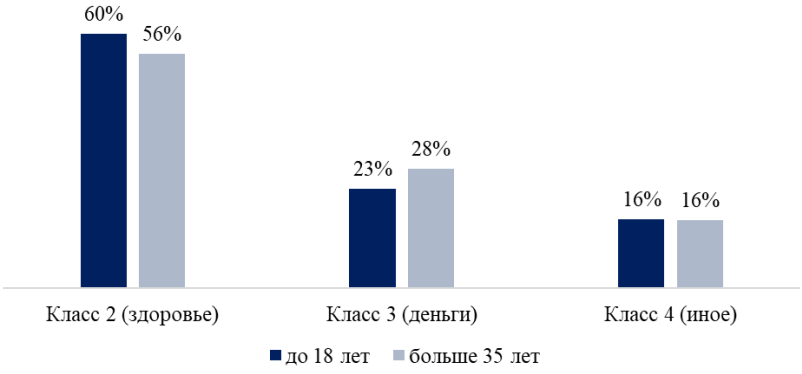
3 класс (от 35 лет и выше) — 430 строк.

Мы определяли три возрастные группы: до 18 лет, 19—34 года, старше 35 лет. Из-за нехватки набора данных в определенных возрастных группах было принято решение использовать технологию для искусственного наращивания данных — SMOTE.

С использованием датасета «ВКонтакте» была выполнена классификация комментариев при помощи LSTM-модели.

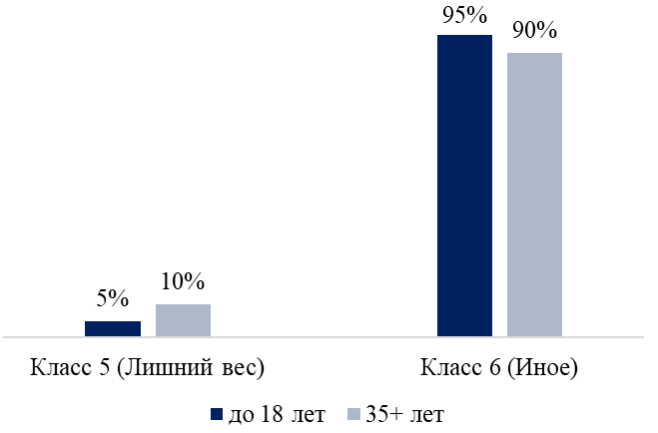
Точность определения возрастной группы составила соответственно 49%, 76% и 53%. Несмотря на недостаточно удовлетворительный результат, разработанный алгоритм мы все же апробировали на нашей базе по курению. Результаты нашей классификации комментариев по возрасту представлены на рисунках 8 и 9. Комментариев, относящихся к классу 2 (19—34 лет), оказалось 35, поэтому по ним данные на рисунках не представлены (см. таблицу 2 Приложения). Существенных различий среди полученных нами возрастных групп не выявлено[5]. Молодежь чуть реже пишет о деньгах и лишнем весе, чаще — о здоровье.

Рис. 8. Результаты классификации комментариев по типу довода бросить курить и по возрасту (до 18 лет и 35+ лет), %



[5] Поскольку классификация текстов по возрасту авторов на данный момент является сложной задачей с большим количеством ограничений, расширенный датасет был классифицирован по возрасту в сотрудничестве с коллегами из Томского государственного университета систем управления и радиоэлектроники (ТУСУР) [Соболев и др., 2022] по альтернативной методике коллег, вычленяющей подростковую группу (см. [Kalabikhina, Kazbekova, Moshkin, 2025]). Отметим, что различий по возрастным группам при использовании классификации коллег из ТУСУР также не наблюдается.

Рис. 9. Результаты классификации комментариев
по типу довода не бросать курить и по возрасту (до 18 лет и 35+ лет), %



Заключение и дискуссия

Нам удалось классифицировать наличие доводов и типы доводов (не)отказа от курения табака. Альтернативный подход к классификации одновременно всех указанных классов, исполненный в данной работе, снижает точность классификации по сравнению с вычленением каждого довода отдельно [Калабихина и др., 2024]. Мы связываем снижение точности классификации с низким наполнением некоторых классов. Однако такой подход более продуктивен для мониторинга самосохранительного поведения. Предполагаем, что увеличение массива данных существенно улучшит результаты точности классификации при одновременном подходе, поскольку наполнение отдельных классов будет достаточным.

Разработан алгоритм классификации комментариев одновременно по полу анонимизированного комментатора и по типам доводов с достаточной точностью с помощью генеративного искусственного интеллекта. Существенных различий в доводах отказа от курения не найдено, женщины чуть больше заботятся о здоровье, мужчины — о бюджете. Похожий вывод можно сделать и в отношении доводов против отказа от курения. Аргумент о страхе набрать вес характерен не только для женщин (как это было по результатам наших эконометрических исследований), но и для мужчин, хотя и в несколько меньшей степени.

Комментарии классифицированы также по возрасту автора. Точность классификации была ниже либо не удавалось классифицировать стандартные укрупненные группы. Существенных различий в полученных возрастных группах по доводам отказа от курения табака не выявлено (молодежь чуть реже отмечает деньги и лишний вес).

Ограничения и перспективы исследования

Ограничения исследования выявлены в трех аспектах. Это (1) недостаточное наполнение комментариями отдельных классов, что снижает точность; (2) отсут-

ствие учета пересекающихся доводов; (3) отсутствие учета контекста (возможного «давления» содержания видеоролика на комментарии под ним). Уточним второй пункт: в рамках текущего подхода не учитываются возможные пересечения, когда в одном комментарии содержится несколько доводов. На прошлом этапе проекта мы проводили эксперименты отдельно для каждого класса, соответственно, пересечения были учтены. Среди размеченных нами на прошлом этапе проекта комментариев доля содержащих несколько доводов составляла 9%.

Ограничение исследований с использованием данных социальных сетей носят общий характер, тем не менее мы кратко обсудим их здесь. Во-первых, это вопрос репрезентативности данных. Например, выборка пользователей YouTube по нашей теме смещена в пользу мужчин и молодежи; она не репрезентует генеральную совокупность — население России.

Во-вторых, специфика больших данных — их неструктурированная природа. Результат в определенной степени зависит от подходов к структурированию данных авторов исследования, что надо принимать во внимание. Собранные и используемые нами данные — это обсуждения пользователей YouTube под видео по теме курения, которые мы не структурируем заранее. Авторам комментариев не задают прямой вопрос о том, почему они хотят или не хотят бросить курить. Мы вводим предпосылку, что если они упоминают тот или иной довод в своем комментарии, то он для них наиболее значим. Это одно из важных ограничений анализа — если в конкретном комментарии автор указывает лишь один довод, это не значит, что другой довод для него/нее не важен.

В-третьих, наличие ботов. Мы предпринимаем процедуру очистки от ботов. Однако надо понимать, что найденные комментарии примерно в объеме 50% являются ботами, несмотря на борьбу организаторов социальной сети с этим явлением. Это увеличивает усилия по набору баз данных.

В-четвертых, возможно искажение информации о поле/возрасте авторами комментариев. В нашей теме это не является ограничением исследования, поскольку еще не введено наказание за нарушение возрастных и иных ограничений при участии в беседах на темы о вредных привычках. И само содержание дискуссии не предполагает мотивов искажений. Но в такого рода исследованиях надо об этом помнить.

В-пятых, современные ограничения генеративных моделей искусственного интеллекта (галлюцинации, неустойчивость результатов и т. д.), которые сегодня активно обсуждаются, могут влиять и на процедуры классификации (в меньшей степени, чем на генерацию текста, однако устойчивость результатов нуждается в дополнительных проверках на разных базах данных).

Отдельно стоит сказать о неустойчивом доступе к разным социальным сетям. Какие-то сети блокируют, какие-то замедляют. Например, замедление работы YouTube в России. В период проведения исследования, начиная с лета 2024 г., в России существенно ухудшилась работа видеохостинга YouTube. Если ситуация сохранится, то в будущем мы ожидаем, что обсуждения по теме курения русскоязычных пользователей переместятся на альтернативные платформы. Возможно, это будут Rutube и «VK Видео», которые являются российскими аналогами YouTube. Но на каждой платформе своя специфика сбора данных и своя

структура пользователей по разным характеристикам, что ограничивает сравнительный анализ.

Перспективы исследования мы видим в первую очередь в совершенствовании классификации доводов самосохранительного поведения (в том числе в вопросах отказа от курения) на укрупненных массивах комментариев социальных сетей; в учете пересечений в доводах; продолжении совершенствования алгоритмов определения пола, возраста, уровня образования комментаторов.

Разработанный нами автоматический алгоритм определения наличия довода по поводу отказа от курения и автоматической классификации доводов по указанным классам можно применять с целью мониторинга мнений пользователей русскоязычных социальных сетей по вопросам отказа от курения. В области рекомендаций для антитабачной политики на основе полученных результатов можно предложить следующие направления, которые могут повлиять на снижение потребления табака:

1) усилить просветительскую работу о вреде табака для здоровья людей, здоровья их детей и других окружающих людей, поскольку этот довод узнаваем и является наиболее популярным;

2) в рамках поощрительных мер развить тезис о сбережении семейного бюджета и времени, а также об альтернативных способах использования сэкономленных средств и минут;

3) бороться с мифами об отсутствии вреда от табака и тем более о наличии пользы;

4) пояснять населению, особенно с учетом различий по полу, что набор веса может наблюдаться при отказе от табака в случае длительного стажа курения, и предлагать реабилитационные программы для сокращения вероятности таких последствий.

Список литературы (References)

1. Калабихина И. Е., Казбекова З. Г., Банин Е. П., Клименко Г. А. Демографические ценности и социально-демографический портрет пользователей ВКонтакте: есть ли связь? // Вестник Московского университета. Серия 6. Экономика. 2023. № 3. С. 157—180. https://doi.org/10.55959/MSU0130-0105-6-58-3-8.
Kalabikhina I. E., Kazbekova Z. G., Banin E. P., Klimenko G. A. (2023) Demographic Values and Socio-Demographic Profile of the VKontakte Users: Is There a Connection? *Lomonosov Economics Journal*. Vol. 58. No. 3. P. 157—180. https://doi.org/10.55959/MSU0130-0105-6-58-3-8. (In Russ.)

2. Калабихина И. Е., Казбекова З. Г., Зубова Е. А. Доводы пользователей социальных медиа по поводу отказа от табакокурения (на основе методов машинного обучения) // Вопросы управления. 2024. Т. 18. № 5. С. 48—67. https://doi.org/10.22394/2304-3369-2024-5-48-67.
Kalabikhina I. E., Kazbekova Z. G. Zubova E. A. (2024). Arguments of Social Media Users Regarding Smoking Cessation (Machine Learning-Based Data). *Management Issues*. Vol. 18. No. 5. P. 48—67. https://doi.org/10.22394/2304-3369-2024-5-58-67. (In Russ.)

3. Кузнецова П. О. Почему не снижается курение у женщин: результаты микроанализа // Женщина в российском обществе. 2019. № 3. С. 91—101.
Kuznetsova P. O. (2019) Why the Number of Smoking Women Does not Decrease: A View from Microanalysis Level. *Woman in Russian Society*. No. 3. P. 91—101. (In Russ.)

4. Сбоев А. Г., Рыбка Р. Б., Молошников И. А., Наумов А. В., Селиванов А. А. Сравнение точностей методов на основе языковых и графовых нейросетевых моделей для определения признаков авторского профиля по текстам на русском языке // Вестник НИЯУ МИФИ. 2023. Т. 10. № 6. С. 529—539. https://doi.org/10.56304/S2304487X21060109.
Sboev A. G, Rybka R. B., Moloshnikov I. A., Naumov A. V., Selivanov A. A. (2023). Comparison of the Accuracies of Methods Based on Language and Graph Neural Network Models for Determining Author Profile Features from Russian Texts. *Vestnik Nacional'nogo Issledovatel'skogo Yadernogo Universiteta "MIFI"*. Vol. 10. No. 6. P. 529—539. https://doi.org/10.1134/S2304487X21060109. (In Russ.)

5. Соболев А. А., Федотова А. М., Куртукова А. В., Романов А. С., Шелупанов А. А. Методика определения возраста автора текста на основе метрик удобочитаемости и лексического разнообразия // Доклады Томского государственного университета систем управления и радиоэлектроники. 2022. Т. 25. № 2. С. 45—52.
Sobolev A. A., Fedotova A. M., Kurtukova A. V., Romanov A. S., Shelupanov A. A. (2022) Methodology to Determine the Age of the Text's Author Based on Readability and Lexical Diversity Metrics. *Proceedings of TUSUR University.* Vol. 25. No. 2. P. 45—52. (In Russ.)

6. Bickel W. K., Tomlinson D. C., Craft W. H., Ma M., Dwyer C. L., Yeh Y. H., Tegge A. N., Freitas-Lemos R., Athamneh L. N. (2023) Predictors of Smoking Cessation Outcomes Identified by Machine Learning: A Systematic Review. *Addict Neuroscience*. Vol. 6. Art. 100068. https://doi.org/10.1016/j.addicn.2023.100068.

7. Cheng N., Chandramouli R., Subbalakshmi K. P. (2011) Author Gender Identification from Text. *Digital Investigation.* Vol. 8. No. 1. P. 78—88. https://doi.org/10.1016/j.diin.2011.04.002.

8. Chu K.-H., Colditz J., Malik M, Yates T., Primack B. (2019) Identifying Key Target Audiences for Public Health Campaigns: Leveraging Machine Learning in the Case of Hookah Tobacco Smoking. *Journal of Medical Internet Research*. Vol. 21. No. 7. Art. e12443. http://dx.doi.org/10.2196/12443.

9. Coughlin L. N., Tegge A. N., Sheffer C. E., Bickel W. K. (2020). A Machine-Learning Approach to Predicting Smoking Cessation Treatment Outcomes. *Nicotine & Tobacco Research: Official Journal of the Society for Research.* Vol. 22. No. 3. P. 415—422. https://doi.org/10.1093/ntr/nty259.

10. Culotta A. (2010) Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: *Proceedings of the First Workshop on Social Media Analytics*. New York, NY: Association for Computing Machinery P. 115—122. https://doi.org/10.1145/1964858.1964874.

11. Dieleman L. A., van Peet P. G., Vos H. M. M. (2021) Gender Differences within the Barriers to Smoking Cessation and the Preferences for Interventions in Primary Care a Qualitative Study Using Focus Groups in The Hague. *BMJ Open.* Vol. 11. No. 1. Art. e042623. https://doi.org/10.1136/bmjopen-2020-042623.

12. Guida M., Otmakhova Y., Hovy E., Frermann L. (2025) LLMs for Argument Mining: Detection, Extraction, and Relationship Classification of Pre-Defined Arguments in Online Comments. *arXiv*. Preprint arXiv:2505.22956. https://doi.org/10.48550/arXiv.2505.22956.

13. Himdi H., Shaalan K. (2024) Advancing Author Gender Identification in Modern Standard Arabic with Innovative Deep Learning and Textual Feature Techniques. *Information*. Vol. 15. No. 12. Art. 779. https://doi.org/10.3390/info15120779.

14. Kalabikhina I., Zubova E., Loukachevitch N., Kolotusha A., Kazbekova Z., Banin E., Klimenko, G. (2023). Identifying Reproductive Behavior Arguments in Social Media Content Users' Opinions through Natural Language Processing Techniques. *Population and Economics*. Vol. 7. No. 2. P. 40—59. https://doi.org/10.3897/popecon.7.e97064.

15. Kalabikhina I., Kazbekova Z., Moshkin V. (2025) (Non)Smoking Comments Classified by Arguments, Genderand Age [Data Set]. *Zenodo*. January 31. Version v1. https://doi.org/10.5281/zenodo.14782953.

16. Kavuluru R., Sabbir A. K. M. (2016) Toward Automated E-Cigarette Surveillance: Spotting E-Cigarette Proponents on Twitter. *Journal of Biomedical Informatics*. Vol. 61. P. 19—26. http://dx.doi.org/10.1016/j.jbi.2016.03.006.

17. Kim K., Kim S. (2025) Large Language Models' Accuracy in Emulating Human Experts' Evaluation of Public Sentiments about Heated Tobacco Products on Social Media: Evaluation Study. *Journal of Medical Internet Research*. Vol. 27. Art. e63631. https://doi.org/10.2196/63631.

18. Klein A. Z., Magge A., Gonzalez-Hernandez G. (2022) ReportAGE: Automatically Extracting the Exact Age of Twitter Users Based on Self-Reports in Tweets. *PloS One*. Vol. 17. No. 1. Art. e0262087. https://doi.org/10.1371/journal.pone.0262087.

19. O'Connor K., Golder S., Weissenbacher D., Klein A. Z., Magge A., Gonzalez-Hernandez, G. (2024) Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review. *Journal of Medical Internet Research*. Vol. 26. Art. e47923. https://doi.org/10.2196/47923.

20. Ritchie H., Roser M. (2023, November) Smoking. *Our World in Data*. URL: https://ourworldindata.org/smoking (date of access: 20.08.2025).

21. Romanov A. S., Kurtukova A. V., Sobolev A. A., Shelupanov A. A., Fedotova A. M. (2020) Determining the Age of the Author of the Text Based on Deep Neural Network Models. *Information*. Vol. 11. No. 12. Art. 589. https://doi.org/10.3390/info11120589.

22. Sboev A., Litvinova T., Gudovskikh D., Rybka R., Moloshnikov I. (2016) Machine Learning Models of Text Categorization by Author Gender Using Topic-Independent Features. *Procedia Computer Science*. Vol. 101. P. 135—142. https://doi.org/10.1016/j.procs.2016.11.017.

23. Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. (2018) Automatic Gender Identification of Author of Russian Text by Machine Learning and Neural Net Algorithms in Case of Gender Deception. *Procedia Computer Science*. Vol. 123. P. 417—423. https://doi.org/10.1016/j.procs.2018.01.064.

24. Visweswaran S., Colditz J. B., O'Halloran P. H., N. R., Taneja S. B., Welling J., Chu K. H., Sidani J. E., Primack B. A. (2020) Machine Learning Classifiers for Twitter Surveillance of Vaping: Comparative Machine Learning Study. *Journal of Medical Internet Research*. Vol. 22. No. 8. Art. e17478. http://dx.doi.org/10.2196/17478.

25. Younkin V., Litvak M., Rabaev I. (2024) Automatic Gender Identification from Text. *Applied Sciences.* Vol. 14. No. 24. Art. 12041. https://doi.org/10.3390/app142412041.

Приложение

Таблица 1. ***Структура причин (не)отказа от курения в разрезе пола (число комментариев каждого класса, шт.)***

Класс	Женщины	Мужчины
Доводы бросить курить		
Класс 2 (Здоровье)	430	681
Класс 3 (Деньги)	129	320
Класс 4 (Иное)	164	191
Доводы не бросать курить		
Класс 5 (Лишний вес)	95	104
Класс 6 (Иное)	673	1567

Источник: составлено авторами на основе обработки комментариев пользователей YouTube с использованием больших языковых моделей LLM (определение типа довода и пола).

Таблица 2. ***Структура причин (не)отказа от курения в разрезе возрастной группы (число комментариев каждого класса, шт.)***

Класс	Автору текста меньше 18 лет	Автору текста больше 35 лет
Доводы бросить курить		
Класс 2 (Здоровье)	659	638
Класс 3 (Деньги)	256	325
Класс 4 (Иное)	177	185
Доводы не бросать курить		
Класс 5 (Лишний вес)	51	216
Класс 6 (Иное)	990	1885

Источник: составлено авторами на основе обработки комментариев пользователей YouTube с использованием больших языковых моделей LSTM (определение типа довода и возраста).