ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

*На правах рукописи*

Кольцова Елена Юрьевна

# ПРИМЕНЕНИЕ МЕТОДОВ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЯЗЫКА ДЛЯ ИССЛЕДОВАНИЯ ОСВЕЩЕНИЯ МЕЖЭТНИЧЕСКИХ ОТНОШЕНИЙ И ДРУГИХ СОЦИАЛЬНО-ПРОБЛЕМНЫХ ТЕМ В БОЛЬШИХ МАССИВАХ ПОЛЬЗОВАТЕЛЬСКИХ ТЕКСТОВ

**Резюме**
диссертации на соискание ученой степени
доктора филологических наук

Москва 2024

# Общие сведения о работе

## Научная проблема исследования и ее актуальность

Распространение интернета и, в особенности, сервисов для коммуникации между пользователями, не являющимися профессиональными медиа-производителями, создает большое количество цифровых следов, в том числе – текстовых. Эти следы являются богатым материалом для исследователей из социальных наук, так как содержат большое количество ценной информации об обществе, его элементах и процессах. Изучение таких следов в социологических целях требует развития новых лингвистических методов, непривычных для социальных исследователей, но также нетипичных и для лингвистов. Учитывая большой объем данных, эти методы в основе своей опираются на автоматическую обработку естественного языка и достижения компьютерных наук, для которых, однако, решение социологических задач также не находится в центре внимания. Всё это создает дефицит релевантных методологий, в особенности целостных методологических подходов, содержащих все этапы: перевод социологической задачи на язык компьютерной лингвистики, построение выборок текстов, удовлетворяющих социологическим запросам, автоматическую обработку текстов и совмещение их с «ручным» контент-анализом для получения результатов, релевантных для социальных наук.

Настоящая диссертационная работа состоит из восьми статей, опубликованных в рамках реализации единого проекта, и посвящена результатам разработки методологий анализа больших объемов текстов пользователей интернета в области межэтнических отношений. Межэтнические отношения – одна из традиционных проблематик в социологии, до недавнего времени изучавшаяся, в основном, с помощью опросных методов. В последнее время, однако, внимание социологов, политологов, психологов приковано к процессам в сфере межэтнических отношений, происходящих в интернете, в особенности к дискуссиям на эту тему, часто выливающихся в речь ненависти. Исследования показывают, что социальные медиа вносят свой вклад в «оффлайновые» межэтнические конфликты и преступления на почве межэтнической ненависти (Williams et al 2020, Chang et al 2016). Некоторые исследования также показывают позитивное влияние межэтнических онлайн-дружб на межэтническую кооперацию и снижение межэтнической напряженности как за рубежом (Žeželj et al 2017), так и в России (Galyapina and Molodikova 2023).

Это определяет важность изучения того, как рядовые пользователи социальных медиа репрезентируют представителей различных этнических групп, включая свою; в частности каково их отношение к разным группам, эксплицитно выраженное в текстах, и каковы контексты их описания, указывающие на имплицитные стереотипы (например, «кавказцы» могут чаще всего появляться в контексте преступности). Получение ответов на эти вопросы требует развития двух типов методов автоматического анализа текстов пользователей. Методы машинного обучения (МО) без учителя, такие как тематическое моделирование, полезны для выявления заранее не известных контекстов упоминания этнических групп, в то время как методы обучения с учителем, такие как нейросетевые алгоритмы классификации, решают задачу автоматического обнаружения в текстах заранее определенных типов репрезентаций этнических групп и их взаимоотношений. Эти два семейства алгоритмов дополняют друг друга, а их соединение может позволить ответить на вопросы о том, какие типы репрезентаций и каких именно этнических групп преобладают в тех или иных контекстах и, соответственно, могут помочь содержательно интерпретировать природу негативных репрезентаций

(например, репрезентация одних групп как опасных может оказаться связанной с религиозным контекстом, а других – с контекстом антисанитарии и риска распространения инфекций).

Адаптация указанных алгоритмов к таким задачам и встраивание их в аналитический инструментарий социологов в виде готовых методологических цепочек имеет большое практическое значение. Так, раннее обнаружение всплесков конфликтных дискуссий в социальных сетях может помочь предотвращению конфликтов в «оффлайне», либо сгладить их последствия, а привлечение внимания к текстам с описанием позитивного межэтнического взаимодействия может помочь наладить межэтнический диалог.

**Некоторые термины**

- Этническая группа – социальная группа, идентифицируемая внешними и внутренними наблюдателями по нескольким из следующих признаков: язык, культура, религия, ареал проживания, фенотипические черты, «кровь». В социологии описывается как социальный конструкт (нечеткое и, возможно, «нереальное» понятие, реально используемое в обществе и имеющее социальные последствия). Не представляет собой термин, определенный экспертами из социальных наук.
- Репрезентация этнической группы – способ речевого описания группы или ее взаимоотношений с другими группами, создающий определенный образ или стереотип ее восприятия.
- Эксплицитная репрезентация этнической группы – выражение отношения к группе или ее взаимоотношениям с другими, которое может быть понято читателем без обращения к другим текстам; например, репрезентация этнической группы как высшей или низшей, опасной или безопасной, либо отношений как конфликтных или мирных.
- Имплицитная репрезентация – избирательное упоминание группы в определенных контекстах, которое, как правило, можно проследить только путем анализа большого количества текстов (например, упоминание центрально-азиатских этнических групп в контексте стройки, рынков и инфекций, а французов – в контексте высокой моды, курортной жизни и путешествий).
- Речь ненависти – в данной работе: негативный полюс общего отношения к этнической группе или персонажу в тексте, где общее отношение – тип эксплицитной репрезентации.
- Текст – одно или ряд высказываний одного и того же автора, размещенные им как отдельная публикация (статья, пост, комментарий).
- Этноним – слово или группа слов, обозначающая в тексте этническую группу или ее представителя, включая некоторые квазиэтнонимы (кавказцы) и этнофолизмы (этнонимы с унизительными коннотациями).
- Этническая группа как именованная сущность (ЭГИС) – совокупность всех этнонимов, описывающих одну и ту же этническую группу, которую можно описать одним «корневым» этнонимом.

**Цель и задачи исследования**

**Цель исследования** – разработать, протестировать и увязать в комплексную методологию новые модификации и способы практического применения автоматического

анализа языка для социологических задач (с фокусом на анализ репрезентаций этнических групп и межэтнических отношений в текстах социальных медиа).

**Задачи**

1. Сформулировать социологическое операционализируемое определение понятия «репрезентации этнической группы в речи» и ее типов.
2. Разработать комплексную и верифицируемую методику ручной разметки текстов пользователей социальных медиа по выделенным типам репрезентаций и методику разметки результатов тематического моделирования.
3. Создать набор корпусов текстов социальных медиа, с разметкой и без, репрезентирующих разные срезы пользователей.
4. Определить применимость тематического моделирования для вычленения контекстов репрезентации этнических групп, для поиска наиболее этнорелевантных текстов с целью последующего ручного анализа, а также для сравнения выраженности различных этнорелевантных тем между собой.
5. Протестировать вариации тематического моделирования с частичным обучением для более эффективного вычленения узких (в данном случае этнорелевантных) тем.
6. Разработать и протестировать доменно-специфичный сентиментный словарь для социально-политических текстов социальных медиа.
7. Протестировать широкий спектр классификационных моделей и способов предобработки текстовых данных, в том числе с использованием сентиментного словаря, для задач выявления различных типов эксплицитных репрезентаций этнических групп.

**Положения, выносимые на защиту**

1. Предложенная типология репрезентаций этнических групп в текстах социальных медиа позволяет дифференцировать этнические группы между собой по характеру их стереотипизации.
2. Классическое тематическое моделирование (LDA без модификаций) вычленяет этнорелевантные темы как пример редких тем из коллекций среднего размера, прошедших предварительный отбор по релевантности, и не решает эту задачу на случайных и на очень больших выборках текстов. Это не позволяет использовать его как инструмент для измерения представленности редких тем в общем дискурсе социальных медиа, но позволяет сравнивать представленность разных этнорелевантных тем между собой в меньших выборках, репрезентирующих специальных подмножества пользователей.
3. Алгоритм ISLDA, в котором группы целевых этнонимов закрепляются каждая за своим диапазоном тем, позволяет вычленять больше этнорелевантных тем (особенно в решениях с большим количеством тем), в среднем более концентрированных и когерентных, чем классический LDA. ISLDA подходит для изучения тем об ограниченном количестве этнических групп на коллекциях среднего размера с относительно длинными текстами.
4. Алгоритм семейства BigARTM, регуляризированный специальным образом и принимающий словарь этнонимов как отдельную модальность, позволяет находить этнорелевантные интерпретируемые темы лучше, чем без словаря. Такой алгоритм подходит для очень больших коллекций относительно длинных текстов, предварительно отобранных по словарю этнонимов.

5. BigARTM с расширенным словарем позволяет вычленять больше тем, в среднем несколько более этноерелевантных и несколько менее интерпретируемых, чем BigARTM с органиченным словарем этнонимов. Это позволяет рекомендовать использовать для частичного обучения тематического моделирования расширенный словарь, содержащий, кроме этнонимов, этнические прилагательные и названия стран.
6. Разработанный сентиментный словарь PolSentiLex опережает RuSentiLex в выявлении негативного сентимента в этнорелевантных текстах, позитивного и негативного сентимента в социально-политических блогах и не отличается в качестве выявления позитивного сентимента в этнорелевантных текстах.
7. Классы репрезентаций уровня текста предсказываются с достаточным уровнем качества при единице анализа «текст» (наличие межэтнического конфликта, позитивного межэтнического взаимодействия, общий негативный и позитивный сентимент)
8. Репрезентации отдельных этнических групп, в частности, речь ненависти по отношению к отдельным группам, предсказывается при единице анализа «ЭГИС», при этом качество трехклассовой классификации (негативная, или речь ненависти, позитивная и нейтральная репрезентации) превышает качество двухклассовой классификации (наличие либо отсутствие речи ненависти).
9. Нейросетевые алгоритмы существенно превосходят классические классификаторы в задаче выявления речи ненависти, даже несмотря на относительно небольшой размер размеченной коллекции.
10. Предсказание речи ненависти на уровне ЭГИС с помощью искусственных нейросетей демонстрирует лучшее качество, чем предсказание на уровне текста, что оказывается осуществимым путем добавления этнонима, обозначающего целевую ЭГИС, в качестве парного текста к тексту, в котором она встречается, при формулировании задачи как парной классификации перед подачей в модель BERT и далее в классификатор.
11. Лингвистические признаки улучшают качество предсказания в совокупности, но не по отдельности, приводя к наилучшим результатам (а) среди классических классификаторов – в архитектуре voting classifier, а (б) среди нейросетевых алгоритмов в комбинации с Conversational RuBERT и последующим плотным слоем.

Данные положения далее упоминаются в формулировке типа «положение i» в тех местах раздела «Содержание работы», где речь идет о получении результатов, соответствующих тому или иному положению.

**Новизна научного исследования**

Общая научная новизна данного междисциплинарного исследования заключается в следующем:

- Впервые сформулировано развернутое определение понятия «репрезентация этничности в тексте» и его типы, протестированные с точки зрения их способности дифференцировать этнические группы.
- Впервые эмпирически определены настройки методов тематического моделирования, оптимальные для выявления этнорелевантных интерпретируемых тем и, соответственно, имплицитных репрезентаций этничности в больших коллекциях текстов социальных медиа.
- Впервые предложен высокоточный алгоритм выявления общего эксплицитного отношения к этническим группам в текстах, часть из которых содержит более одной

ЭГИС;

- Впервые создан комплекс коллекций текстов социальных медиа на русском языке, часть из которых имеет большой размер и репрезентирует широкие слои пользователей интернета, а другая часть содержит уникальную разметку по ряду оснований, включая репрезентации этничности.

Помимо общей новизны, работа обладает существенной новизной, специфичной для области лингвистики:

- Предложен комплексный метод оценки качества тематического моделирования, включающий ручную разметку;
- Разработана детализированная методика разметки текстов по набору теоретически обоснованных типов отношений к этническим группам;
- Создан словарь этнонимов, который может использоваться как для отбора этнорелевантных текстов, так и для улучшения качества работы тематического моделировании при выявлении этнорелевантных тем;
- Создан доменно-специфичный словарь для опознавания тональности общественно-политических текстов в социальных медиа. Показана его положительная роль в работе классификаторов при опознавании этнической речи ненависти;
- Показана положительная роль других лингвистических признаков для улучшения качества алгоритмов, опознающих речь ненависти;
- Впервые проведен качественный анализ ошибок работы нейросетевых алгоритмов и выделены лингвистические признаки текстов, облегчающие и затрудняющие опознавание этнической речи ненависти;
- Созданные коллекции с разметкой свободно доступны и вносят вклад в уменьшение дефицита лингвистических ресурсов на русском языке;
- Все задачи решались для русского языка впервые.

**Статьи, выносимые на защиту**

1. Bodrunova S., Koltsova O., Koltcov., S., Nikolenko. Who's Bad? Attitudes Toward Resettlers From the Post-Soviet South Versus Other Nations in the Russian Blogosphere // *International Journal of Communication*. 2017. Vol. 11. P. 3242-3264. (Scopus Q1, WoS Q3, Список А)
2. Nikolenko S., Koltcov S., Koltsova, O. Topic modelling for qualitative studies // *Journal of Information Science*. 2017. Vol. 43. No. 1. P. 88-102. (WoS Q2, Scopus Q1, Список А)
3. Apishev M., Koltsov S., Koltsova O. Mining ethnic content online with additively regularized topic models // *Computacion y Sistemas*. 2016. Vol. 20. No. 3. P. 387-403. (Scopus Q3, Список С)
4. Koltsova O. Yu, Alexeeva S. V., Koltcov S. N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media, in: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва,1–4 июля 2016 г.) / Под общ. ред.: В. Селегей. Вып. 15. М. : Изд-во РГГУ, 2016. P. 277-287. Scopus Procs
5. Koltsova O., Alexeeva S., Pashakhin S., Koltsov S. PolSentiLex: Sentiment Detection in Socio-political Discussions on Russian Social Media, in: *Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science* Book 1292:

Communications in Computer and Information Science. Cham: Springer, 2020. P. 1-16. Scopus Procs

6. Koltsova O., Nikolenko S. I., Alexeeva S. V., Nagornyy O. S., Koltsov S. Detecting interethnic relations with the data from social media, in: *Digital Transformation & Global Society: Second International Conference, DTGS 2017, St. Petersburg, Russia, June 21-23, 2017, Revised Selected Papers*. Springer, 2017. P. 16-30. Scopus Procs
7. Koltsova O., Alexeeva S., Nikolenko, S., Koltsov M. Measuring Prejudice and Ethnic Tensions in User-Generated Content // *Annual Review of CyberTherapy and Telemedicine*. 2017 (Scopus Q4, Список С)
8. Pronoza E., Panicheva P., Koltsova O., Rosso, P. Detecting ethnicity-targeted hate speech in Russian social media texts // *Information Processing and Management*. 2021. Vol. 58. No. 6. Article 102674. (WoS Q1, Список А)

**Другие статьи автора по теме исследования**

9. Koltsova O., Pashakhin S. Agenda Divergence in A Developing Conflict: A Quantitative Evidence from A Ukrainian and A Russian TV Newsfeeds // *Media, War and Conflict*. 2020. Vol. 13. No. 3. P. 237-257. (Scopus Q2, Список А)
10. Koltsova O., Koltsov S., Nikolenko S. I. Communities of co-commenting in the Russian LiveJournal and their topical coherence // Internet Research. 2016. Vol. 26. No. 3. P. 710-732. ( WoS Q1, Scopus Q1, Список А)
11. Nagornyy, O., & Koltsova, O. Y. (2019) Redefining media agendas: topic problematization in online reader comments. *Media and Communication*, Vo.7, No.3, P. 145–156. (Scopus Q2, Список С)
12. Panicheva P., Mararitsa L., Sorokin S., Koltsova O., Rosso P. Predicting subjective well-being in a high-risk sample of Russian mental health app users // *EPJ Data Science*. 2022. Vol. 11. Article 21 (Q1 WoS & Scopus, Список А)
13. Koltsova O., Scherbak A. N. 'LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012 // *New Media and Society*. 2015. Vol. 17. No. 10. P. 1715-1732. (WoS & Scopus Q1, Список А)
14. Koltsova O., Koltcov S. Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal // *Policy & Internet*. 2013. Vol. 5. No. 2. P. 207-227. (Scopus No Q in 2013, Список А)
15. Koltcov S., Ignatenko V. and Koltsova O., Estimating Topic Modeling Performance with Sharma–Mittal Entropy, *Entropy* 2019, 21(7), 660. (WoS Q2, Scopus Q2, Список С)
16. Koltsova O. Social media for joint experimental, survey and observational data collection: The case of VKontakte (VK). The SAGE Handbook of Social Media Research Methods, 2nd ed., 2022. (WoS-indexed handbook)
17. Кольцова Е. Ю., Маслинский К. А. Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: методология, методы, математическое моделирование. 2013. № 36. С. 113-139. (Список D)

Данная нумерация статей используется далее для ссылок на них при описании содержания работы. Нумерация статей 1-8 дана в порядке выполнения работ (не в порядке публикации).

**Финансирование**

Выполнение работ по данному исследованию поддержано следующими грантами:

- Грант РНФ № 15-18-00091 «Разработка концепции и методологии многоуровневого мониторинга состояния межнациональных отношений по данным социальных сетей».
- Грант РГНФ № 14-04-12031 «Разработка общедоступной базы данных и краудсорсингового веб-ресурса для создания инструментов сентимент-анализа».
- Гранты ПФИ НИУ ВШЭ: 83-2014, 68-2017, 67-2018, 61-2019 (частично).

## Содержание работы

### Данные

В исследовании использованы специально созданные для него коллекции текстов двух типов: развернутые тексты блогов, представляющие собой либо оригинальные посты, либо репосты профессиональных медиа-текстов, написанные литературным языком, и короткие тексты (посты и комментарии) пользователей социальных сетей, написанные неформальным языком.

1. LJposts: коллекция представляет собой примерно 1.58 млн текстов за период один год (с середины 2013 по середину 2014) и включает все посты топ-2000 блоггеров платформы Живой Журнал (LiveJournal), самой популярной платформы для общественно-политического блоггинга в России в то время. Закачка данных производилась еженедельно согласно текущему рейтингу блоггеров. Коллекция использована целиком в статье 3 для тематического моделирования на больших данных и в статьях 1, 2, 4 для формирования подмножеств 1.1-1.5.
    - 1.1. LJposts11: подмножество LJposts, включающее 11 недель с 4 февраля по 19 мая 2013 года и 363 579 постов. Период отобран по социологическим критериям. Использована в статье 1 для конструирования выборки LJposts11ethnic.
    - 1.2. LJposts11ethnic: размеченное подмножество LJposts11 из 990 текстов, куда вошли топ 30 текстов из каждой из 33 тем тематического решения, признанных этнорелевантными. Разметка произведена по ранней версии кодировального листа и фиксирует разные типы репрезентаций этнических групп (в т.ч. низший / высший, свой / чужой и др.)
    - 1.3. LJPosts12: подмножество LJposts, включающее 4 первых месяца 2013 года и 235 407 постов. Использована в статье 2 для апробации алгоритма ISLDA.
    - 1.4. LJpostsRandom4: четыре коллекции, представляющих 4 случайных месяца из 2012 и 2013 годов, сформированные по принципу LJPosts. Использованы в статье 2 для апробации метрики качества tf-idf-когерентность.
    - 1.5. LJpostsPol: подмножество LJposts из 70 710 постов, куда вошли тексты, имеющие вероятность не менее 0.1 по 104 темам тематического решения, признанным социально-политическими. Используется в статьях 4, 5 для формирования первой версии сентиментного словаря.

2. LJcomments: Все комментарии к постам коллекции LJposts, где все комментарии к одному посту агрегированы в один текст. Около 0.9 млн текстов. Использована в статье 5 для формирования коллекции LJcommentsPol.
   2.1. LJcommentsPol: подмножество LJcomments, сформированное аналогично коллекции LJpostsPol. Включает 15 188 агрегированных текстов комментариев, имеющих вероятность не менее 0,001 по 88 темам тематического решения, признанным социально-политическими. Использована в статье 5 наряду с LJpostsPol для формирования окончательной версии сентиментного словаря.

3. VKrandom: коллекция всех текстов со стен 74 303 пользователей ВКонтакте, случайно отобранных из каждого субъекта федерации РФ пропорционально региональной аудитории ВКонтакте. Всего 9 168 353 постов и 933 516 комментариев к ним; период сбора: лето 2015. Использована в промежуточных экспериментах для статей 3, 6.

4. RuEthnics: коллекция всех сообщений (постов и комментариев) из всех русскоязычных социальных медиа за 1.5 года (январь 2014 – декабрь 2015), содержащих хотя один из 115 пост-советских ЭГИС из авторского словаря этнонимов. Охват всех социальных медиа и сбор первичных данных обеспечивался компанией IQBuzz. После дедупликации и другой предобработки коллекция содержит 2 660 222 текстов о 97 этнических группах. Использована целиком в статье 8 для дообучения алгоритма ConversRuBERT и в статьях 6, 7, 8 для формирования подмножеств 4.1-4.3.

   4.1. RuEthnicsMarked1: подмножество RuEthnics из 7181 текстов, отобранных так, чтобы представлять все 97 этнических групп, и размеченных по окончательной версии кодировального листа, фиксирующего разные типы репрезентаций этничеких групп (в т.ч. низший / высший, агрессор / жертва, опасный / неопасный) и межэтнических отношений (в т.ч. есть ли этнический конфликт). Использовалась в статье 6 для пилотного обучения алгоритма распознавания различных типов репрезентаций.
   4.2. RuEthnicsMarked2: подмножество RuEthnics и расширение RuEthnicsMarked1. Содержит 14998 текстов, размеченных по окончательной версии кодировального листа и некоторые новые данные по добавленным текстам. Использовалась в статье 7 для тестирования влияния размера коллекции на качество классификации.
   4.3. RuEthnoHate: подмножество RuEthnics и расширение RuEthnicsMarked2, которое включает процедуру отбора наиболее качественно размеченных текстов из RuEthnicsMarked2 и процедуру досэмплинга и доразметки. Итоговая коллекция содержит 5 594 текста и используется в статье 8 для обучения нейросетевого классификатора предсказанию общего отношения к этнической группе (негативное = речь ненависти, нейтральное и позитивное).

Датасеты 1.2 и 4.1.-4.3. имеют двухуровневую структуру, где есть переменные уровня текста (например, дата) и уровня ЭГИС (отношение к ней). Наблюдением в этих датасетах является совокупность упоминаний ЭГИС в данном тексте; таким образом, на каждый текст может приходиться разное количество наблюдений в датасете (по числу упомянутых в нем этнических групп). Эти датасеты размечались тремя независимыми кодировщиками.

Обоснование построения корпусов. Живой Журнал был выбран как самая популярная на момент сбора данных блог-платформа в России социально-политического содержания. В ходе исследования было обнаружено, что наиболее популярные блоггеры имеют больше

сходств с профессиональными медиа-производителями, чем с обычными интернет-пользователями, а середина рейтинга зашумлена ботами. Поэтому позднее было принято решение о построении выборки пользователей ВКонтакте – социальной сети, ставшей к моменту следующего сбора самой популярной. Использовалась случайная выборка, стратифицированная по субъектам РФ пропорционально представленности регионов в аудитории ВКонтакте; неактивные аккаунты отсекались. Данная выборка полностью удовлетворяет критериям социологической репрезентативности (аудитории ВКонтакте), однако на ней были получены в основном негативные результаты. Поэтому было принято решение о построении сходной выборки, но обогащенной этнорелевантной информацией. Выборка RuEthnics сформирована как сплошная выборка этнорелевантных сообщений за определенный период, что также близко к социологическим критериям репрезентативности, за исключением того, что в ней, возможно, не представлены этнорелевантные тексты, не упоминающие этнонимов. В целом, группа корпусов RuEthnics представляет собой разумный компромисс между социологическими потребностями и возможностями методов автоматической обработки языка.

**Типы репрезентаций этнических групп в текстах и инструмент их измерения**

Теоретически обоснованное и четко операционализированное определение того феномена, классы которого предстоит выявить с помощью машинного обучения, является залогом успеха классификации, в особенности для таких нечетких категорий, с которыми приходится работать социологам. Теории межэтнических взаимоотношений и межгруппового взаимодействия в социологии и психологии сходятся в том, что отношение к этнической группе, как своей, так и чужой – это сложный многосоставной феномен, содержащий имплицитную и эксплицитную составляющую. В социологии хорошо развит инструментарий измерения отношения к этническим группам в опросах общественного мнения, а в исследованиях медиа широко представлено направление выявления искажений (biases) в репрезентациях различных социальных групп путем ручного анализа выборок медиа-текстов, зачастую качественного. Однако до сих пор не предложено надежных инструментов фиксации того, как отношение к этническим группам и его аспекты репрезентированы в публичных текстах пользователей социальных медиа – то есть, в ситуациях, когда рядовые граждане не отвечают на структурированный опросник и когда анализируемый текст не является результатом профессионального медиа-производства.

На основе анализа различных теорий в данной работе предложен инструмент, состоящий из следующих вопросов, предназначенных для специально обученных разметчиков текстов.

***Вопросы уровня текста***

А. Упоминается ли в тексте межэтнический конфликт?

Б. Упоминается ли в тексте позитивное межэтническое взаимодействие?

***Вопросы уровня ЭГИС:***

В. Говорит ли автор об этнической группе в целом или о конкретном персонаже?

Г. Каково общее отношение автора в этнической группе или персонажу?

Д. Принадлежит ли автор к этнической группе, о которой говорит?

Е. Описывается ли группа или персонаж как низшая или высшая по сравнению с другими?

Ж. Описывается ли данная группа или персонаж как жертва или агрессор в межэтнических отношениях?

З. Описывается ли данная группа или персонаж как опасная?

И. Призывает ли автор к насилию по отношению к данной группе?

Теории стереотипизации указывают на то, что для мнений людей характерны сверх-обобщения, когда определенные черты приписываются недифференцированно группе в целом, и это находит отражение в вопросе В. Теории внутригруппового смещения (ingroup bias) говорят о том, что люди склонны видеть свою группу в более позитивном свете, соответственно, репрезентация группы как своей (себя как принадлежащего группе) должна быть связана с ее более позитивным образом (вопрос Д). Теории речи ненависти обращают внимание на необходимость дифференцировать разные ее степени, что находит отражение в вопросах Е, З и И. Наконец, следует отметить, что большинство теорий сфокусировано на негативном полюсе измерения отношения к этническим группам, а данное исследование ставило своей задачей выявление отношений обоих полюсов. Это находит свое отражение во всех двуполярных вопросах (Г, Д, Ж) и в паре вопросов А и Б. Кроме того, вопрос А нацелен на выявление социально опасного содержания даже в тех текстах, где отношения к конкретным группам трудно определить, но конфликтность прослеживается (например, в текстах, где автор агрессивно защищает свою этническую группу от нападок оппонента неизвестной национальности).

**Словарь этнонимов**

Словарь этнонимов применялся на разных этапах работы и был создан по следующей схеме: 1. Наименования этнических групп были собраны из данных ООН, Российской переписи населения и некоторых других источников, без разделения этнических («араб») и национальных («иракец») групп. 2. Из них был сформирован список пост-советских этнических групп, который был дополнен мета-этнонимами («славянин»), некоторыми регионализмами («кавказец»), квази-этнонимом «казак» и этнофолизмами, список которых формировался экспертным путем. Подробно принципы формирования списка описаны в статье 6. 3. Список был автоматически расширен производными («армянка», «армяшка») и релевантными биграммами («армянская девочка», «армянский народ»). Все производные, включая этнофолизмы (например, «жид» и его производные) были сгруппированы под названием этнической группы («еврей»), за исключением этнофолизмов, не указывающих ни на одну этническую группы однозначно («хач», «чурка»), которые были выделены в отдельные группы производных («хачиха», «хачонок»). В базовом списке пост-советских этнонимов 115 таких групп.

**Общие принципы работы методов машинного обучения, использованных в работе**

**А. Обучение без учителя: тематическое моделирование**

Тематическое моделирование (ТМ) – группа методов, которые тем или иным образом производят сжатие размерности матрицы слов на документы, где строка – документ (текст), столбец – слово, ячейка – абсолютная или взвешенная частота слова в тексте. ТМ по своей функции сходно с нечеткой ко-кластеризацией столбцов и строк такой матрицы, в результате чего и тексты, и слова оказываются сгруппированы в заданное число групп. В ТМ эти группы понимаются как латентные переменные с приписанным смыслом «тема», а задачей является восстановление скрытых распределений слов и документов по темам по наблюдаемы данным – распределению слов по документам. Будучи вероятностным алгоритмом, ТМ приписывает

каждый текст и каждое слово к каждой теме, но с разной вероятностью. Выходными данными являются две матрицы – матрица вероятностей слов в темах и матрица вероятностей тем в документах. Пользователь может отсортировать как слова, так и тексты в этих матрицах по вероятности их принадлежности к i-той теме; при качественном ТМ наиболее вероятностные слова, как правило, дают представление о содержании темы, а наиболее вероятностные тексты дают представление о дискурсе, характерном для данной темы. Таким образом, ТМ дает пользователю возможность быстро оценить тематическую структуру большой коллекции, не поддающейся чтению, ничего не зная о темах заранее, и сконцентрироваться на чтении только самых релевантных для задач пользователя документов. Наиболее распространенные версии ТМ – группа алгоритмов Latent Dirichlet Allocation (LDA) и более ранний pLSA (probabilistic Latent Sematic Analysis).

ТМ имеет ряд нерешенных проблем, включая отсутствие надежных и общепринятых метрик качества и, соответственно, критериев выбора параметров алгоритмов. Проблемами, с которыми пришлось иметь дело в данном исследовании, являются: (1) недостаток способности выявлять темы, если тексты, где эти темы сильно выражены, составляют очень малую долю в коллекции, (2) слабая способность к масштабированию при переходе от коллекций среднего размера (порядка сотен тысяч документов) к очень большим коллекциям (порядка миллионов документов) и (3) неспособность работать с короткими текстами. Для проблем 1 и 2 в данном исследовании предложены решения, а для проблемы 3 предложены способы ее косвенного сглаживания. ТМ поддается регуляризации – процедуре добавления новой информации, ограничивающей поиск решений и подталкивающих его в определенном направлении. Это может быть максимизация различия тем (декорреляция) или фиксация определенных слов / групп слов за определенными темами / группами тем (частичное обучение). Эти возможности были использованы в данном исследовании (этапы 1, 2).

**Б. Обучение с учителем: классификаторы**

Классификация, предназначенная для разделения объектов (например, текстов) на заранее определенные классы, предполагает следующую процедуру. Алгоритм получает информацию о наблюдениях, их признаках и принадлежности к классам (обучающая коллекция), после чего в функции, предсказывающей класс как зависимую переменную, итеративно подбирает коэффициенты либо для каждого признака-предиктора, либо иным образом так, чтобы классы предсказывались максимально правильно по заданной мере качества. Затем такая обученная модель с фиксированными коэффициентами запускается для предсказания классов на коллекции, на которой алгоритм не обучался (тестовая коллекция). По этому тестовому запуску рассчитываются меры качества работы алгоритма, среди которых самые распространенные precision, recall, F1-мера и accuracy. Как правило, вся процедура повторяется несколько раз с разным разбиением имеющихся размеченных данных на тестовую и обучающую коллекции (кросс-валидация). Примером простейшего классификатора является логистическая регрессия.

Искусственные нейронные сети (ИНС) – это сложные комплексы функций и алгоритмов, общий принцип работы которых отдаленно сходен с биологическими нейросетями. Как и простые классификаторы, ИНС итеративно подбирают веса признаков в модели предсказания классов, начиная со случайных или одинаковых. Эти веса подаются для преобразования сразу в несколько функций, называемых искусственными нейронами или нодами нейросети, в простом случае – один признак на один нод. Группа нодов, одновременно получающих сигнал о всех весах, называется слоем, а весь нейросетевой алгоритм может содержать от одного до нескольких слоев, где каждый последующий слой получает

информацию о весах, скорректированную на предыдущем слое. В конце предсказание, основанное на весах признаков, подобранных системой, сравнивается с правильным ответом, и процесс повторяется итеративно тем или иным способом.

Как правило, базовая структура ИНС дополняется рядом дополнительных методов, таких как: функция активации (порог веса признака, ниже которого нод обнуляет данный вес и посылает далее нулевой сигнал), внимание (алгоритм, дифференцирующий признаки по важности), исключение (алгоритм случайного исключения обученных нейронов из дальнейшего использования с целью предотвращения переобучения модели), конкатенация (соединение групп признаков, полученных разными способами) и др., которые были использованы в и в данной работе. Нейросетевые алгоритмы также способны самостоятельно генерировать признаки из сырых данных и принимать на вход дополнительную информацию, которой нет в данных – например, вместо слов принимать векторные представления слов (ВПС).

ВПС – это результат работы отдельного алгоритма ИНС, который обучался не на исследуемой коллекции, а на других данных большего объема. Простейшее векторное представление слова – это представление его в виде последовательности (вектора) чисел, соответствующих частотам его совместной встречаемости со всеми словами коллекции. На практике используются вектора сжатой размерности, где измерениями являются не слова, а абстрактные латентные переменные, являющиеся результатом сокращения размерности пространства слов. Поскольку ВПС содержат информацию о совместной встречаемости слов в гораздо больших массивах текстов, чем изучаемые, они позволяют ИНС лучше «понимать» тексты, особенно если в них мало информации – например, если тексты очень короткие, а потому пересечение словарей даже похожих по смыслу текстов минимально (что мешает отнесению их к одному классу). Отметим, что если ВПС обучены на текстах, не похожих на изучаемую коллекцию, они могут привносить в модель нерелевантную информацию, но для таких случаев разработаны подходы дообучения ВПС на изучаемой коллекции. Для каждой классификационной задачи требуется настройка всех перечисленных аспектов работы алгоритма, что и было сделано на этапе 4 данного исследования.

**Этап 1: пилотирование инструментария измерения репрезентаций**

На данном этапе были поставлены следующие задачи: (а) применить стандартное тематическое моделирование для выявления этнорелевантных текстов и оценить его эффективность с помощью асессоров; (б) также с помощью асессоров выделить в подмножестве наиболее релевантных текстов типы контекстов и типы эксплицитных репрезентаций этнических групп; (в) провести социологический анализ методами стандартной статистики, чтобы определить, какие типы репрезентаций и какие контексты характерны для каких этнических групп, а также позволяет ли разработанный инструментарий дифференцировать между этническими группами и находить осмысленные зависимости.

Отбор текстов производился из коллекции LJposts11, а ручной разметке подверглась коллекция LJposts11Ethnic. На этих коллекциях поставленные задачи были решены успешно (статья 1). После экспериментов с разным числом тем, был предложен оптимум (300-400 тем на коллекции такого рода текстов порядка $10^4$ документов). В тематическом решении на 300 тем было выявлено 33 этнорелевантных и интерпретируемых темы, в некоторых из которых контексты восприятия этнических групп были ярко выражены. Примеры 20 наиболее вероятных слов в таких темах включают: «олень , ханты , праздник , чум , сургут , север , оленевод , стадо , сургутский , тундра , рог , место , упряжка , якутия , мансийск , снегоход , охотник»; «мигрант, чечня, республика, грозный, чеченский, страна, российский, гражданин,

рамзан, кадыр, кадыров, чеченец, гражданство, миграционный, азия, депардье, режим». Экспертное чтение наиболее вероятных текстов по этим темам выявило, что ханты и манси в них ассоциируются с традиционным промыслом и сводятся к нему, а чеченцы отождествляются с иностранными мигрантами, имеющими меньше прав на пребывание на принимающей территории, чем коренные жители. Ручная разметка на контексты (политический, экономический, социальный, культурный) также показала статистически значимую дифференциацию между укрупненными группами этничностей (табл. 5, 6 статьи 1). Так, в освещении центрально-азиатских групп преобладали социальный и экономический контексты, южно- и северо-кавказских – политический, причем для южно-кавказских был также существенен культурный контекст, а для других коренных народов РФ – только культурный.

Далее было показано, что укрупненные этнические группы хорошо дифференцируются шкалами Г, Е, Ж «общее негативное / позитивное отношение», «жертва – агрессор», «высший – низший» - например, северо-кавказские этнические группы репрезентировались как значимо более агрессивные, чем южные (табл.3 статьи 1). Также хорошую дифференцирующую силу показала шкала «свой – чужой» (табл. 4 статьи 1), однако асессоры столкнулись с трудностями ее определения в большом числе текстов, поэтому в более поздней версии она была заменена на вопрос об этнической принадлежности автора, если она очевидна (вопрос Д). Вопросы, спрашивающие о наделении репрезентируемых групп действиями и прямой или косвенной речью, не показали большой полезности для дифференциации между группами и далее были исключены из инструмента измерения репрезентаций. Вопросы З, И собрали очень мало данных, чтобы делать выводы об их полезности, поэтому они были оставлены в инструменте для дальнейшего изучения. Таким образом, был сформирован инструмент измерения репрезентаций этничности, имеющий статистически доказанную дифференцирующую силу (положение 1, выносимое на защиту).

Кроме того, было выявлено, что пост-советские этнические группы сильно уступают по уровню общего внимания нациям, имеющим глобальное или региональное значение (американцы, немцы), но последние репрезентируются не как этнические группы, а как представители своих стран и, соответственно, в контексте не межэтничеких отношений, а международных. Это заставило нас выдвинуть гипотезу о необходимости дифференциации этих контекстов для более эффективного выявления именно межэтнических отношений, что было проверено на следующем этапе исследования.

**Этап 2: настройка и тестирование тематических моделей для выявления имплицитных контекстов репрезентации межэтнических отношений**

Первой была опробована идея закрепления заданного набора слов за рядом, или интервалом тем, то есть обнуление вероятностей слов из этого набора во всех темах тематического решения, кроме входящих в заданный интервал. Этот подход получил название ISLDA – interval semi-supervised Latent Dirichlet Allocation – и сначала был апробирован в нескольких экспериментах на датасете LJPosts11. Были использованы интервалы по 3 темы для 22 наиболее употребляемых в коллекции этнических групп (т.е. всего 66 целевых тем в решении на 200 тем) и проведена проверка качества в форме эксперимента на асессорах по методу подмешивания слова (word intrusion) и аналогичному ему методу подмешивания темы (topic intrusion), оценивающих связность тем и способность алгоритма приписывать тексты к релевантным темам, соответственно. Эксперименты показали более высокое качество целевых тем в ISLDA по описанным метрикам по сравнению с нецелевыми темами и по сравнению со всеми темами обычного LDA (табл. 4 статьи 2). Кроме того, целевые темы ISLDA оказались

лучше аналогичных тем в LDA и по метрике tf-idf-когерентности (табл. 5 статьи 2), а эксперименты с числом тем показали, что преимущество ISLDA перед LDA в умении находить больше целевых тем по данной этничности проявляется при большем числе тем в решении. Эти результаты отражены в положении 3. Метрика tf-idf-когерентности была также предложена в статье 2 и апробирована на коллекции LJPosts12, показав лучшие результаты, чем традиционная когерентность.

Далее идея частичного обучения была развита следующим образом: было предложено закрепить весь словарь этнонимов за широким диапазоном тем, который составил бы 30-50% от общего числа тем в тематическом решении. Этот подход решает проблему выбора размера интервала в ISLDA и существенно расширяет число используемых этнонимов. Подход решено было проверить на базе группы алгоритмов, основанных на pLSA и реализованных в библиотеке BigARTM. Эта библиотека обладает рядом преимуществ, включая простое встраивание ряда регуляризаторов, а также функционал параллельных вычислений на основе деления коллекции на порции (батчи), что позволило использовать его на большой коллекции LJPosts в 1.5 млн постов. Одной из подзадач стала проверка того, как деление коллекции на бачти влияет на качество результатов. Всего было обучено 8 разных моделей на 400 тем; из них первые две были базовыми (baseline): (1) pLSA и (2) LDA в имплементации BigARTM; вторые две (3, 4) делили темы на фоновые (250 тем) и целевые (150) таким образом, что появление этнонимов в фоновых темах пенализировалось, а в целевых поощрялось, также был использован ряд других регуляризаторов; модели 5, 6, помимо предыдущих приемов, использовали словари как отдельную модальность: в модели 5 был использован словарь этнонимов; в модели 6 он был дополнен этническим прилагательными и названиями стран / провинций в случае отсутствия этнонимов в коллекции (грузин -> грузинский, Грузия). Так проверялась гипотеза о необходимости разделения тем о межэтнических и международных отношениях. Модели 7, 8, помимо всего предыдущего, использовали разные виды рекурсии: на вход следующего этапа моделирования подавались тексты, отобранные на предыдущих этапах.

Качество моделей оценивалось как по когерентности и tf-idf-когерентности, так и множеством асессоров, которым было предложено оценить понятность принципа объедения слов в темы, содержательную интепретируемость темы и ее релевантность межэтническим и международным отношениям. По сравнению с экспериментами по подмешиванию слова, такой инструмент позволяет различить разные причины низкого качества тем. По всем метрикам оценки качества модели 5, 6 заметно опередили все остальные (табл. 2, 4, 5, 6 статьи 3), несущественно отличаясь друг от друга по разным аспектам (Положение 4). Важно отметить, что добавление в словарь стран и прилагательных заставило модель выявлять больше тем по международным отношениям, однако не подорвало качество и количество тем по межэтническим отношениям, поэтому модель 6 была рекомендована как ведущая к использованию (Положение 5). Кроме того, общее качество тем, несмотря на деление коллекции на батчи, по tf-idf когерентности оказалось лучше, чем у ISLDA, а по экспертным оценкам – не хуже.

Далее эксперименты с BigARTM были продолжены на коллекции VKrandom – наилучшей из всех с точки зрения социологической репрезентативности. Все эти эксперименты, включая тесты на текстах, объединенных по автору, дали негативные результаты, не позволив выделить интерпетируемые и / или релевантные темы. Тогда эксперименты были перенесены на коллекцию RuEthnics, которая уступает VKrandom тем, что в ней могут быть упущены этнорелевантные тексты, не содержащие этнонимов. На RuEthnics был расширен диапазон настраиваемых параметров, включая число тем, но лучшее качество показала базовая модель pLSA, что можно объяснить тем, что роль частичного обучения выполнил предварительный отбор текстов по наличию в них этнонимов. Более успешное

моделирование на RuEthnics по сравнению с VKrandom, вероятно, можно объяснить наличием в RuEthnics определенной доли длинных текстов, сфокусированных на какой-либо одной этнорелевантной теме; вероятно, удлиннение текстов ВК за счет механического объединения всех постов одного автора в один не облегчило задач тематического моделирования потому, что такие квази-тексты не приобрели никакой доминирующей тем или небольшой группы доминирующих тем.

Из данной серии экспериментов можно сделать следующие выводы (отражающие положения 2-5). Задача выявления этнолеревантных тем из очень больших коллекций пользовательских текстов требует предварительной фильтрации таких коллекций. Если в приоритете задача выявления неочевидных контекстов репрезентации этничности – то есть тем, включающих тексты, в которых может не быть этнонимов, - коллекцию не следует фильтровать по ключевым словам, но следует сформировать из относительно длинных текстов и использовать модель 6 bigARTM. Если в приоритете выявление тем в коротких текстах, где с большей вероятностью содержатся спонтанные реакции пользователей, коллекцию не следует формировать только из длинных текстов, при этом отфильтровав по наличию этнонимов. Если такая выборка будет представительной с точки зрения разных социальных медиа (что предпочтительно), она с большой долей вероятности будет содержать некоторую долю длинных текстов, и это позволит тематическому моделированию быть успешным даже при использовании моделей без частичного обучения (но с параллелизацией вычислений).

При оценке качества получающихся тематических моделей следует обращать внимание на следующие: (1) насколько темы различны между собой (большая доля повторяющихся тем – признак сбоя или неоптимальных настроек; (2) насколько списки наиболее вероятных слов соответствуют спискам наиболее вероятных текстов (несоответствие – признак некачественной темы); (3) насколько темы интерпретируемы по наиболее вероятным словам; (4) насколько темы интерпретируемы по наиболее вероятным текстам. Быстрая автоматическая оценка качества тем по словам возможна с помощью метрики tf-df-когерентности как наиболее соответствующей человеческой разметке. При ручной оценке опора на прямые вопросы, в отличие от опоры на эксперименты, дает возможность различить темы, не имеющие содержания, но с понятным принципом формирования (например, темы, сформированные вокруг обсценной лексики или вокруг названий месяцев), и темы, имеющие собственное содержание, такие как приведенные выше темы о хантах и о чеченцах.

**Этап 3: разработка сентиментного словаря**

Сентиментный словарь, получивший название PolSentiLex, был разработан для целого ряда задач, включая определение общего сентимента социально-политических пользовательских текстов и вклад в формирование признакового пространства для алгоритмов, обучаемых классифицировать эксплицитные репрезентации этнических групп. Основная идея при составлении словаря для таких задач заключалась в том, что сентиментный вес должен присваиваться словам исходя из того значения, которое они имеют именно в контексте социально-политических пользовательских текстов. Для этого была использована следующая методика (статья 4).

Была сформирована коллекция LJPosts, так как среди топовых блоггеров Живого Журнала на тот момент доминировали авторы острых социально-политических текстов. На основе ТМ на 200 тем была сформирована коллекция LJPostPol (социально-политические тексты), а также выделены слова-кандидаты для сентиментного словаря, по 200 наиболее вероятностных слов из каждой из тем, размеченных как социально-политические. Эти слова была дополнены кандидатами из существовавших словарей (словаря RuSentiLex Н.Лукашевич

на тот момент не существовало), составив первую версию прото-словаря из 9,539 слов, 7,546 из которых встретились в коллекции LJPostPol. Далее был создан интерфейс, в котором асессорам каждое слово предъявлялось в тексте (было подобрано по три разных текста на слово) и предлагалось оценить общий сентимент слова в данном контексте и общей сентимент текста по шкале от -2 до 2. Всего было получено по 32,437 меток слов и текстов, которые были откодированы не менее чем тремя независимыми асессорами, чьи оценки были усреднены и округлены. Качество словаря было оценено путем предсказания получившихся сентиментых оценок текстов на основе словаря, поданного в ПО SentiStrength. Значения мер precision и recall оказалось выше случайного и на уровне существовавших предсказаний для русского языка на сходных задачах (табл. 3 статьи 4), но не достаточно высоко, чтобы считать работу законченной.

В продолжение данной работы были сфомированы коллекции LJComments и LJCommentsPol, и вся описанная выше процедура была повторена на текстах комментариев (статья 5). Вторая часть прото-словаря составила 9 539 слов, из которых 6 860 встретилось в коллекции LJCommentsPol; было получено 26 851 оценок. Окончательная версия словаря, полученная после объединения двух частей, включила 3 924 слова, признанных не-нейтральными. К моменту окончания работ был выпущен словарь RuSentiLex, и оценка качества PolSentiLex проводилась в сравнении с ним. Работа обоих словарей тестировалась на коллекции LJPostsPol, на котором PolSentiLex обучался, и на коллекции RuEthnicsMarked2, на котором он не обучался. Тексты последней содержат две отдельные метки общего негативного и общего позитивного сентимента по трехбалльной шкале (не путать с общим отношением к этнической группе); в тестах предсказывался именно общий сентимент. Сравнению подверглись: оба словаря, встроенные в SentiStrength (словарный подход), и оба словаря, использованные как наборы признаков в лучших решениях трех классических алгоритмов машинного обучения: SVM, KNN и NB. В результате (а) словарный подход существенно опередил машинное обучение на обеих коллекциях; (б) PolsentiLex показал сходное качество с RuSentiLex в моделях с машинным обучением и существенное преимущество при словарном подходе, как описано в положении 6; (в) негативный класс предсказывался лучше, чем позитивный (рис. 1, 2, 3 статьи 5).

Резюмируя этот этап, можно сказать, что преимущество PolSentiLex перед RuSentiLex в данной задаче не удивительно, потому что PolSentiLex – доменно-специфичный словарь, а RuSentiLex – общий, и они не заменяют друг друга. Более важным результатом оказалась способность простого словарного подхода давать качество выше 60% по разным метрикам в задаче на три класса: это убедило нас, что наш словарь содержит ценную информацию, которая может оказаться полезной в алгоритмах МО следующего поколения (ИНС), поэтому он был взят в дальнейшую работу.

**Этап 4: тестирование алгоритмов МО для предсказания эксплицитных репрезентаций этнических групп**

В исходном эксперименте этого этапа предсказывались репрезентации А, Б (на два класса), Г (на три класса – позитивный, негативный, нейтральный) и общий негативный и позитивный сентимент текста на коллекции RuEthnicsMarked1, которая была создана первой (статья 6). Коллекция формировалась так, чтобы этнические группы были представлены в пропорциях, в которых они представлены в коллекции RuEthnics, с перепредставленностью ряда самых редких групп. Для первого эксперимента все оценки асессоров были усреднены, а строкой был сделан текст, таким образом, по переменной Г (общее отношение к этнической группе) предсказывалось отношение, доминирующее в тексте и не дифференцированное по

отдельным ЭГИС. В качестве классификатора была использована простая логистическая регрессия, а в качестве признаков – 6 364 наиболее частотных слов и 7 039 наиболее частотных биграмм. Качество получившихся предсказаний (табл. 1 статьи 6) получилось заметно выше случайного, но все же довольно низким по главной переменной Г и требующим дальнейшего совершенствования. По переменным Д, Е, Ж, З, И в датасете оказалось недостаточно данных. Поэтому далее коллекция была расширена и вошла в состав большего датасета RuEthnicsMarked2, сформированного по тому же принципу. Для предсказания переменных уровня текста (А, Б и сентимента) строкой в датасете по-прежнему являлся текст. Для предсказания переменной Г строкой была сделана пара ЭГИС-текст (для текста i, где упоминались этнонимы x, y, формировались строки типа i-x, i-y). Поскольку тело текста для строк i-x и i-y одинаково, для различения между ними при формировании матрицы признаков в признаковое пространство слов и биграмм была добавлена информация о той ЭГИС, в отношение которой должно быть предсказано значение переменной Г по заданной строке. Процедура классификации была повторена. Результаты (табл. 1, 2 статьи 7) показывают улучшение качества по всем переменным и классам, кроме метрики recall по общему позитивному сентименту. Для recall по негативному классу общего отношения и для precision по негативному и позитивному классам общего отношения улучшение было существенным (15-18%). Эти результаты отражают положения 7 и 8.

Поскольку объем данных по переменным Д-И увеличился несущественно, было принято решение сосредоточиться на переменной Г, главной для наших задач, и провести масштабные эксперименты с большим набором классических классификаторов и нейронных сетей (статья 8). Перед обучением была сформирована более качественная коллекция RuEthnoHate, с обогащенным негативным классом и очищенная от текстов с низкой согласованностью кодировщиков. Предсказание осуществлялось по переменной Г: на три класса (негативный, позитивный, нейтральный) на уровне пары ЭГИС-текст (аналогично статье 7) и на два класса (негативной, другое) на уровне текст. Для предсказания на уровне ЭГИС в ИНС информация о ней была добавлена в форме отдельного текста в архитектуре парных предсказаний (где первый слой ИНС получает информацию не из одного входа, а из двух).

В качестве базовых (baseline) моделей были использованы классические классификаторы (SVM, NB, LR, VC) и ряд наборов признаков, включая униграммы как базовый набор, негативные слова из PolSentiLex, а также (для классификации на уровне ЭГИС) слова с повышающим коэффициентом из контекстного окна вокруг целевого этнонима (полный список см. раздел 4.1. статьи 8). Наборы признаков, модели и их параметры, включая размер окна, перебирались по сетке. Среди этих алгоритмов при предсказании на уровне ЭГИС лучшим оказался VC (voting classifier) со всеми наборами лингвистических признаков (включая PolSentiLex), ни один из которых не улучшал качество по отдельности, но совместно со всеми другими дал значимый прирост (табл. 4 статьи 8) (Положение 11а).

На вход ИНС вместо слов подавались ВПС, включая: модель Word2Vec CBOW, модели Word2Vec, обученные на НКРЯ и на RuEthnics, а также Conversational RuBERT базовый и дообученный на RuEthnics. Были обучены сотни моделей ИНС с разными входными данными, архитектурами и настройками отдельных компонентов. В качестве базовых моделей из числа ИНС были взяты модели LSTM+GRU; схематичное представление об основных апробированных архитектурах этой группы моделей дано на рис. 1 статьи 8. Далее, для Conversational RuBERT был опробован ряд более сложных архитектур, представленных на рис. 2 статьи 8. Важной особенностью двух из трех этих архитектур является то, что наборы признаков, полученные на выходе из BERT, затем конкотирировались (соединялись) с лингвистическим признаками, описанными в разделе 4.1. статьи 8. По результатам всех экспериментов (табл.6(2) статьи 8) лучшей оказалась модель 6, представленная на рис. 2,

панель с. Она сочетает признаки из Conversational RuBERT с лингвистическими признаками, которые совместно подаются в плотный слой, но, в отличие от модели 7, не содержит слоя LSTM (Положение 11б). Ее общая F1=0.892, а F1-мера по целевому классу (негативному) F1hate=0.813. Это выше не только всех моделей, классических (положение 9) и нейросетевых, выполняющих классификацию на три класса на уровне ЭГИС, но даже выше, чем у алгоритмов, выполняющих классификацию на два класса на уровне текста (где лучший результат F1=0.864 и F1hate=0.760) (Положение 10). Поскольку предсказание отношения к этническим группам на уровне ЭГИС производилось впервые не только для русского языка, но и вообще, его невозможно было сравнить с аналогами.

В целом наши эксперименты говорят о следующем. Во-первых, лингвистические признаки важны для улучшения предсказания репрезентаций этничности. Это говорит о перспективности дальнейшей работы над форматами входных данных для нейросетевых алгоритмов. Во-вторых, классификация на уровне ЭГИС (аналог аспектов в ABSA) показывает лучшие результаты, чем классификация на уровне текста. Это подтверждает верность нашего предположения о том, что отношение к этнической группе не выражается на уровне текста в целом и, кроме того, может различаться для разных групп в одном и том же тексте, что не дает возможности алгоритму классифицировать текст в целом. В-третьих, различение трех типов отношений более эффективно, чем поиск только речи ненависти. Это парадоксально, потому что классификационные задачи на два класса, как правило, решаются проще, чем на три. Это свидетельствует о верности нашей стратегии ухода от бинарного понятия речи ненависти (есть / нет) к понятию общего отношения, для которого характерно три, а не два состояния (негативное, позитивное и нейтральное). Такое различение позволяет избегать смешения нейтрального и позитивного классов, что и приводит к более высоким результатам. В-четвертых, важные выводы получены при анализе ошибок работы нейросетевых алгоритмов. Установлено, что все алгоритмы справляются лучше, если отношение к ЭГИС выражено лексикой, попадающей внутрь контекстного окна, и хуже, если оно выражено непрямо (ирония, шутка, использование негативных стереотипов) и с помощью сложной синтаксической структуры (множественные отрицания, вопросы, анафора). Преодоление этих проблем, а также предсказание других типов репрезентаций этничности, не охваченных данным проектом, могут составить предмет для будущих исследований.