Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Санкт-Петербургский государственный университет»

На правах рукописи

Мамаев Иван Дмитриевич

ЛИНГВИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ СКРЫТЫХ СООБЩЕСТВ В КОРПУСЕ СОЦИАЛЬНЫХ МЕДИА С ПРИМЕНЕНИЕМ МУЛЬТИМОДАЛЬНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

Специальность: 5.9.8. Теоретическая, прикладная и сравнительно-сопоставительная лингвистика

Диссертация на соискание ученой степени
кандидата филологических наук

Научный руководитель:
кандидат филологических наук, доцент
Митрофанова Ольга Александровна

Санкт-Петербург
2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В современном мире происходит стремительное развитие информационно-коммуникационных технологий, в частности, социальных сетей, которые стали часто используемым каналом коммуникации говорящих. С точки зрения компьютерной лингвистики каждую социальную сеть можно рассматривать как корпус, а пользовательскую страницу – как подкорпус, состоящий из авторских текстов и репостов. Работая с такими подкорпусами, исследователи выделяют общие параметры, на основании которых обнаруживаются определенные пользовательские сегменты – скрытые сообщества. Изучение скрытых сетевых сообществ имеет ряд важных аспектов, оказывающих положительное влияние на различные области знаний. Например, с точки зрения социологии закрытые форумы и группы в социальных сетях и частные чаты могут предоставить уникальную информацию об общественном мнении и менталитете участников [Градосельская, Щеглова, Карпов, 2019; Binesh, Rezghi, 2018; Gmati et al., 2018]. В криминологии скрытые сетевые сообщества могут служить площадкой для планирования и организации незаконных действий, включая киберпреступления, терроризм, торговлю наркотиками и оружием. Изучение таких сообществ помогает правоохранительным органам выявлять угрозы и противостоять им [Кириченко, Радивилова, Барановский, 2017; Аванесян и др., 2020; Hopkins, 2010]. В психологии анализ дискуссий в сетевых сообществах помогает раскрыть многие психологические аспекты, такие как проявление субъективности, поведенческие модели и воздействие на массовое сознание [Воронин, Ковалева, Чеповский, 2020; Litvinova, Sboev, Panicheva, 2018]. Наконец, изучение сетевых сообществ может помочь предсказывать и предотвращать социальные конфликты [Abaidullah, Ahmed, Ali, 2015; Kamta, Scheffran, 2022; Wong et al., 2022]. В последнее десятилетие скрытые сообщества попали в сферу интересов компьютерных лингвистов: исследователи подробно изучают коэффициенты лексического разнообразия и логической связности [Попов, Чеповский, 2022, с. 41-42], тематические компоненты групп [Lafia et al., 2022] и пр. Лингвистическое исследование скрытых сообществ обуславливается необходимостью разработки

специализированных критериев для создания корпуса текстов, выбора оптимального алгоритма для идентификации тем, объединяющих пользователей, и детализации основных параметров описания сообществ. В этой связи **актуальными** представляются следующие направления:

— разработка корпуса постов социальных сетей русскоязычного сегмента сети Интернет для моделирования скрытых сообществ;

— применение методов искусственного интеллекта для создания оптимального процесса предобработки специализированного корпуса текстов скрытых сообществ;

— отбор языковых признаков для описания разрабатываемых моделей скрытых сообществ;

— лингвостатистическое описание моделей скрытых сообществ.

Степень разработанности темы. До формирования термина *‘скрытые сообщества’* рассматривался ряд смежных понятий. В частности, в работах Дж. Прайса в 1960-х гг. вводился термин *‘невидимые колледжи’*, обозначавший группы реальных ученых, которые публикуют работы на одни и те же темы, но при этом лично не знают друг друга. В последние десятилетия с развитием цифровых коммуникаций акцент сместился в сторону изучения сообществ на онлайн-платформах. Здесь особенно значимыми стали работы по анализу социальных сетей и изучению виртуальных коммуникаций таких авторов, как А.В. Бухановский, Т.А. Литвинова, Д.М. Оболенский, D. Salz и др. Для моделирования структуры скрытых сообществ используются алгоритмы семантической компрессии текстов, в том числе и тематическое моделирование. В исследованиях Н.В. Лукашевич, М.А. Нокель, А.А. Фильченкова, О.А. Митрофановой, Л.В. Тен и др. описываются критерии выбора алгоритма тематического моделирования для отдельных задач, детализируются рекомендуемые параметры настройки процедур и приводится лингвистическая интерпретация полученных тематических моделей. Наконец, особое внимание описанию сетевых структур уделяется в трудах А.А. Чеповского, А.Н. Воронина,

F. Iqbal и др. Однако в этих работах не рассматриваются проблемы описания именно лингвистических профилей скрытых сообществ, выявленных в исследовательских массивах текстов.

Объект исследования – языковые аспекты текстов социальных сетей, представляющих скрытые сообщества. **Предмет** исследования – лингвистические параметры текстов пользователей скрытых сообществ, которые являются основой создания лингвистических профилей.

Гипотеза исследования: тексты тематически аннотированного корпуса социальных медиа, который отобран по определенным критериям и обработан автоматически и вручную, допускают создание модели скрытых сообществ и проведение процедуры лингвистического профилирования, предполагающей выделение лингвостатистических признаков.

Цель настоящего исследования – разработка процедуры лингвистического профилирования модели скрытых сообществ, созданной методами тематического моделирования. Разрабатываемые лингвистические профили будут описывать не речевое поведение отдельных носителей языка, а цифровые проекции групп пользователей социальных сетей, поскольку вопрос о соотношении интернет-данных пользователей и реальных данных остается открытым [Прошкин, 2019; Долгих, Першина, 2021].

Для достижения данной цели необходимо решить следующие **задачи**:

1) обобщить исследовательский опыт по изучению структур сетевых сообществ методами корпусной лингвистики;

2) определить критерии отбора материала для создания корпуса русскоязычных постов социальных сетей с целью моделирования скрытых сообществ;

3) собрать исследовательский корпус по сформулированным критериям;

4) разработать процедуру автоматической предобработки исследовательского корпуса и применить ее к собранному корпусу;

5) устранить ошибки в текстовых данных, которые возникли в результате автоматической предобработки;

6) обосновать выбор алгоритма тематического моделирования как метода семантической компрессии обработанного корпуса и лингвистического метода создания модели скрытых сообществ;

7) детализировать формальную и социальную структуру представленной модели скрытых сообществ;

8) произвести отбор лингвистических параметров, необходимых для описания модели скрытых сообществ;

9) построить профили скрытых сообществ с помощью исследовательской процедуры лингвистического профилирования, которая основана на расчете внутритекстовых коррелятов.

В данной работе применяются **методы** корпусной лингвистики, вероятностного тематического моделирования и комбинаторно-статистических вычислений. В качестве **основного исследовательского инструментария** выступили такие программы как язык программирования *Python*[1], позволяющий за счет привлечения внешних библиотек создавать скрипты различного уровня сложности для обработки корпусов, инструмент *Profiling-UD* [Brunato et al., 2020], с помощью которого извлекаются основные количественные данные о постах пользователей, а также среда *Microsoft Excel*[2] для проведения статистических расчетов.

Теоретической основой настоящей работы послужили как классические, так и современные подходы к исследованию особенностей интернет-текстов [Холодковская, 2014; Crystal, 2001; Squires, 2010; Herring, 2012 и др.], структуры сообществ [Митягин, Якушев, Бухановский, 2012; Мейлахс, Рыков, 2016; Хорошевский, Ефименко, 2017; Bollobas, 1998; Newman, 2003; Fortunato, 2010 и др.], а также разножанровых текстовых коллекций компьютерными методами [Нокель, Лукашевич, 2015; Blei, Ng, Jordan, 2003; Khokhlova, 2017 и др.].

[1] https://www.python.org/
[2] https://www.microsoft.com/ru-ru/microsoft-365/excel

Материалом для создания исследовательского корпуса стал русскоязычный сегмент социальной сети ВКонтакте объемом более 10 000 постов, опубликованных не ранее 01.01.2020.

Достоверность практических результатов работы обеспечивается репрезентативностью эмпирических данных, количественными оценками качества обучения тематических моделей, применением методов оценки согласованности экспертов при разметке тематических моделей, что позволило создать модель скрытых сетевых сообществ, и методов статистических исследований, что позволило представить только значимые корреляты на морфологическом, синтаксическом и лексическом уровнях.

Основные положения, которые выносятся на защиту.

1. Интеграция лингвистических признаков в процесс моделирования скрытых сообществ позволяет обнаружить дополнительные характеристики разрабатываемой модели, которые не учитываются при использовании математических методов. Предлагаемый подход основан не только на количественных показателях разрабатываемой модели, но и на качественных параметрах корпуса, используемого для создания модели.

2. Репрезентативность специализированного корпуса социальных сетей для идентификации скрытых сообществ обеспечивается разработанными критериями отбора данных и их последующей многоступенчатой процедурой фильтрации.

3. Ручное внедрение параметра «автор» в процедуру тематического моделирования позволяет, во-первых, сделать ее мультимодальной, во-вторых, выявить узконаправленные слова-тематизаторы, т.е. единицы, формирующие темы в составе тематической модели.

4. Визуализация созданной модели скрытых сообществ в виде графовой структуры позволяет установить, что плотный центр графа образуют узлы, которые соответствуют пользователям, публикующим тексты в социальных сетях на большое количество тем, в то время как на разреженной периферии находятся узлы, соответствующие пользователям, приверженным одной теме.

5. Процедура лингвистического профилирования, применяемая к исследовательскому корпусу, основывается на сочетании статистических и лингвистических методов анализа. С помощью статистических методов вычисляются различные метрики (например, вычисление средней длины связи зависимости, коэффициента лексической плотности и др.), которые позволяют создать количественную основу для дальнейшего анализа. Ключевым этапом процедуры является вычисление внутритекстовых коррелятов, указывающих на взаимосвязь рассчитанных метрик. Лингвистическая интерпретация сообществ позволяет установить, какие именно языковые единицы и конструкции характерны для определенных групп.

6. Применение методов многомерного анализа итоговых количественных данных обусловлено формой представления профилей скрытых сообществ – кортежем числовых значений.

7. Русскоязычные пользовательские посты в скрытых сообществах наиболее полно характеризуются с точки зрения морфосинтаксических коррелятов при уровне значимости $p < 0.05$.

8. Лексические корреляты текстов пользователей скрытых сообществ практически не являются значимыми при $p < 0.05$, что указывает на лексическую гомогенность постов социальных сетей.

Научная новизна диссертационного исследования состоит в следующем.

1. Разработана процедура создания скрытых сообществ на основе современных семантических анализаторов.

2. Впервые для построения модели скрытых сообществ применяется мультимодальный подход в тематическом моделировании, который, помимо основной триады распределений «*слова – темы − документы*», учитывает и параметр авторства.

3. Впервые введено понятие ‘*лингвистический профиль пользователей скрытого сообщества*’, под которым понимается набор лингвистических

коррелятов, характеризующих особенности построения пользовательских постов с общим тематическим компонентом.

4. Представлена процедура профилирования текстов пользователей как подход лингвистической интерпретации скрытых сообществ в социальных медиа.

Теоретическая значимость исследования заключается в том, что изучение функционирования языка в скрытых сообществах позволяет выявить зафиксированные нормы для участников данных групп. Данное исследование характеризуется междисциплинарным потенциалом, поскольку внедрение лингвистических анализаторов в уже реализованные процедуры выделения групп пользователей позволит более детально описать скрытые сообщества в таких областях знаний, как социология, психология, антропология и др. Результаты исследования вносят вклад в развитие компьютерной лингвистики, корпусной идентификации скрытых сообществ и их лингвистического описания.

Практическая значимость состоит в том, что разработанная методика готова для внедрения в сервисы, обеспечивающие функционирование социальных сетей, например, в системы модерации пользовательских групп, которые учитывают предпочтения авторов постов по ряду лингвистических параметров.

Диссертационное исследование состоит из введения, трех глав, заключения и списка использованной литературы. Во *введении* сформулированы объект, предмет, цель и задачи исследования, а также его научная новизна, актуальность, теоретическая и практическая значимость, выносимые на защиту положения. В *первой главе* рассмотрены сущностные особенности функционирования интернет-текстов. Во *второй главе* раскрыто понятие скрытых сообществ как статического конструкта, который создается на основе больших объемов данных. В этой главе также охарактеризованы алгоритмы моделирования скрытых сообществ – графовые, кластерные и смешанные, при этом последние могут разрабатываться на основе современных инструментов компьютерной лингвистики. В *третьей главе* описываются лингвистические данные, методология эксперимента, приводятся примеры корреляционных расчетов, а также обсуждаются существующие проблемы при работе с данными. В *заключении* обобщены результаты настоящего

диссертационного исследования и намечены перспективы дальнейшей работы. Список использованных источников насчитывает 179 наименований, из них 84 – на русском языке, 95 – на иностранных языках. Использованные данные представлены в репозитории GitHub[3].

Апробация результатов исследования. Основные положения и результаты диссертационной работы были отражены в научных докладах, которые были представлены на научных конференциях российского и международного уровней.

1. Международная конференция Artificial Intelligence and Natural Language Conference (2020, Финляндия, Хельсинки, онлайн).
2. Международный семинар Computational Models in Language and Speech в рамках международной конференции TEL (2020, Россия, Казань, онлайн).
3. XIV Научно-практическая конференция «Инновационные технологии и технические средства специального назначения» (2021, Россия, Санкт-Петербург).
4. 50-я Международная научная филологическая конференция имени Людмилы Алексеевны Вербицкой (2022, Россия, Санкт-Петербург, онлайн).
5. Международный семинар Computational Linguistics в рамках международной конференции Internet and Modern Society (2022, Россия, Санкт-Петербург).

Основные положения и результаты работы отражены в семи публикациях, три из которых – в научных журналах, входящих в Перечень рецензируемых научных изданий Высшей аттестационной комиссии Министерства науки и высшего образования (по научной специальности 5.9.8), а также в изданиях, индексируемых в международной базе данных Scopus.

[3] https://github.com/Wheatley961/Hidden_Communities_Thesis

Глава 1. Веб как корпус: лингвистические особенности функционирования интернет-текстов

Современные интернет-тексты – уникальный пласт языкового материала со специфическими лингвистическими особенностями, обусловленными как техническими возможностями платформ для письменного и устного общения (VK, Skype, Jitsi и пр.), так и коммуникативными потребностями пользователей [Апажева, 2014; Шляховой, 2017]. Совокупность существующих интернет-текстов представляет собой корпус, благодаря которому можно исследовать широкий спектр данных, в том числе клавиатурно-опосредованные и устные тексты в социальных сетях, блогах, форумах, новостных статьях и др. В корпусной лингвистике интернет-тексты рассматриваются в русле направления *WaC* (*Web as Corpus*, *веб как корпус*). Концепция WaC была введена в работах A. Kilgarriff (например, [Kilgarriff, Grefenstette, 2003]) и развивается в ряде отечественных и зарубежных трудов (см., например, [Benko, Zakharov, 2016; Khokhlova, 2017; Kehoe, 2021; Skantsi, Laippala, 2023]). На основании общедоступных интернет-текстов разрабатываются специализированные ресурсы, такие как Генеральный Интернет-Корпус Русского Языка[4], корпуса семейств TenTen[5], Aranea[6] и др. Как при их создании, так и при использовании узконаправленных исследовательских веб-корпусов важно учитывать лингвистические особенности оформления интернет-текстов. Во-первых, данная необходимость связана с употреблением нестандартных языковых конструкций [Гридина, Талашманов, 2019]. Игнорирование этих особенностей приводит не только к снижению качества корпуса, но и качества результатов проводимых исследований. Во-вторых, современные лингвистические процессоры, предназначенные для автоматической обработки текстовых массивов, не учитывают узус социальных медиа, характерные черты которого проявляются в грамматически неверном написании лексических единиц и построении предложений, использовании окказиональной лексики и др.

[4] http://www.webcorpora.ru/
[5] https://www.sketchengine.eu/documentation/tenten-corpora/
[6] http://unesco.uniba.sk/aranea_about/

[Савва, Еременко, Давыдова, 2015; Преминина, 2016]. Перед исследователями стоит задача выбрать оптимальный набор инструментов, которые минимизируют потерю исходных лингвистических данных. Именно поэтому необходимо более подробно рассмотреть особенности функционирования интернет-текстов.

1.1 Графические и орфографические особенности интернет-текстов

При создании интернет-текстов особое внимание уделяется графическим средствам, позволяющим четко и лаконично передать смысл высказывания. Одним из самых распространенных способов выражения эмоций в социальных сетях являются *текстовые смайлики* (*эмотиконы*) – пиктограммы, типографические знаки, изображающие определенную эмоцию [Масликова, 2019, с. 70]. Среди основных функций эмотиконов выделяют такие, как:

1) тонально-ориентированная, с помощью которой осуществляется передача настроения сообщения и которая позволяет установить доброжелательный контакт между собеседниками;

2) ориентирующая, с ее помощью привлекается внимание к обсуждаемому вопросу;

3) разделительная: она позволяет разграничивать как текст, так и эмоциональные блоки, «давая при интенсивном общении время не только на обдумывание ответа, но и на эмоциональную разрядку» [Попова, 2021, с. 436-437].

Денотативное значение смайлика определяется непосредственно при первичном визуальном анализе, однако коннотации, вложенные автором поста в используемый символ, могут остаться нераспознанными другими пользователями: «...только автор знает, какой смысл и значение он вкладывает в текст, сопровождая его тем или иным смайлом» [Смирнова, 2019, с. 78]. В примере «*СПУТНИК V БУДЕШЬ? А БУСТЕРЕНКОЙ ТАКОЙ ВСМ ЧООО Хочу пфайзер* 🗿🗿🗿🗿*и морген такой ля ты летучая мышь* 😂😂😂😂»[78] автор из сообщества «Лентач» использует два смайла: 🗿, который обозначает недоумение с оттенком

[7] Здесь и далее примеры постов социальных сетей приводятся в орфографии пользователей.
[8] https://vk.com/lentach?w=wall-29534144_15827186

недовольства, а также 😂, основной смысл которого – громкий смех [Быкова, 2023]. При этом стоит отметить, что, вполне возможно, автор использует эти эмотиконы, чтобы выразить ироническое отношение к процессу вакцинации.

Наряду с эмотиконами пользователи социальных сетей обращаются к *стикерам*, которые приобретают наибольшую популярность. Стикеры определяют как графические изображения, они «в отличие от смайликов, … имеют конкретных авторов и четко классифицируются на группы. В каждой группе насчитывается от 16 до 48 и более стикеров с различными эмоциями... Главная особенность стикеров заключается в том, что они, как правило, объединены одним героем или одной темой» [Матусевич, 2016, с. 68].

Современные исследования направлены на изучение типов и функций использования стикеров в онлайн-среде. Y. Tang в одной из своих работ [Tang et al., 2021] проанализировала использование стикеров в пяти небольших онлайн-сообществах, в которых азиатские студенты обменивались сообщениями, связанными с учебной жизнью. Использовалось четыре типа стикеров, наиболее востребованной оказалась группа «анимированная картинка без текста» (англ. *animated picture without text*). Было установлено, что функции стикеров делятся на две основные категории: тональная функция сообщения и иллокутивная функция. На основе интервью с семью участниками Y. Tang обнаружила расхождения между намерениями отправителя и интерпретацией получателя для 34.7% стикеров, но эти расхождения не воспринимались коммуникантами как критическая ошибка, поэтому дальнейшее общение проходило без проблем.

В ряде других исследований для изучения смайликов и стикеров применяются методы машинного обучения. Так, в [Al-Maroof et al., 2021] рассматриваются используемые студентами стикеры в приложении WhatsApp. Для решения поставленной задачи авторы интегрировали модель принятия технологий (*technology acceptance model, TAM*) с теорией использования и удовлетворения (*uses and gratifications theory, U&G*). Была разослана анкета для сбора данных 372 студентам различных университетов, которые состояли в учебных групповых

беседах в WhatsApp. На основе применения методов машинного обучения авторы показали, что алгоритм классификации случайный лес (*Random Forest*) превосходит другие классификаторы в предсказывании верного выбора стикера в зависимости от намерений студента с точностью 78.57%. Результаты помогут разработчикам стикеров начать их активное использование в образовательной деятельности.

В работе [Быкова, 2023] проведено исследование, нацеленное на выявление эмоциональной окраски постов и словосочетаний в социальной сети ВКонтакте. Подробно описан процесс получения, обработки и использования набора данных. С помощью методов машинного обучения проведены эксперименты, с использованием метрик качества классификации оценены полученные результаты. Лучший результат по метрике *F1 macro*, равный 69.70%, достигнут с помощью модели *Bag-of-Words + VotingClassifier (soft)* (комбинация «мешка слов» и ансамблевого подхода с мягким голосованием) на нормализованных текстах с пунктуацией и эмодзи. Также были получены лучшие результаты по метрике качества классификации *Weighted F1* – 83.74% – для модели рекуррентной нейросети GRU и 92.92% для дообученной модели на основе *rubert-tiny2*.

Помимо эмотиконов, используется базовый набор знаков препинания и их нестандартные комбинации [Crystal, 2001, p. 89]. Например, неоднократное использование вопросительных знаков свидетельствует об озадаченности пользователя: *«Кому пришло в голову вырвать камеру видеонаблюдения в доме, расположенном по адресу: Кудрово, Европейский пр., д 13, корп. 6? С головой все в порядке???»*[9]. Употребление нескольких восклицательных знаков говорит об эмоциональном порыве пользователя, причем как в положительном аспекте, так и в негативном: *«Ребятыыыы !!! Забирайте железных коней...»*[10].

Стоит отметить, что не только знаки препинания помогают считать определенный тональный элемент сообщения. Важно учитывать и различное

[9] https://vk.com/kudrovolife?w=wall-51766355_2979246

[10] https://vk.com/kudrovolife?w=wall-51766355_2979168

сочетание регистра печатных букв: использование «забористой» комбинации, т.е. последовательное чередование прописных и строчных букв, использование только прописных букв или других сочетаний [Мамаев, 2022b; Herring, 2012]. В нижеприведенном посте пользователь с помощью прописных букв пытается передать иронию: *«ЛДПР хочет, чтоб на ценниках писали не только отпускную, но и закупочную цену. АААА НЕБЕЗУМНАЯ ИНИЦИАТИВА ОТ ЛДПР КОНЕЦ СВЕТА ГРЯДЕТ»*[11].

Наконец, тональная окраска сообщения передается с помощью транслитерации текста сообщения. Существуют стандартные системы транслитерации (например, см. [ГОСТ 7.79-2000]) и нестрогий вариант, в рамках которого общепринятые правила не соблюдаются. Так, М.Н. Крылова приводит фразу *«Ya oru»*: этим способом автор привлекает внимание читателя интересному явлению [Крылова, 2019, с. 131], однако данный формат не поддерживается упомянутым ГОСТом: по его правилам, например, буква *«я»* была бы транслитерирована в виде *«â»*.

В социальных сетях возникают ситуации, когда человеку также необходимо переосмыслить свое мнение по вопросу, что приводит к редактированию исходного сообщения. В этом случае используется преднамеренное зачеркивание текста: «...автор как бы отказывается от первоначального варианта высказывания. Но одновременно он оставляет перечеркнутый вариант в тексте, чтобы показать либо альтернативу мысли, либо ее скрытый подтекст» [Полидовец, 2020, с. 302]. В примере *«Мы с Масей реально ~~кастри~~ саблезубные коты!!!»*[12] автор текста рассказывает о своих домашних питомцах, сопровождая их фотографии фразами, которые могли бы сказать эти самые животные. Один из мяукающих питомцев изначально хотел «сказать», что они кастрированы, но потом передумал и решил сделать акцент на остроте клыков, при этом автором сообщения умышленно оставлено зачеркнутое слово.

[11] https://vk.com/lentach?w=wall-29534144_1403808
[12] https://filibuster60.livejournal.com/1197506.html

Наконец, среди графических средств оформления интернет-текстов выделяют и использование нетрадиционных видов шрифта и комбинаций различных цветов, что приводит к визуальному обособлению текста от остальных и привлекает к нему внимание, а также использование комбинации символов нескольких групп [Крылова, 2016], например, комбинация кириллических и графических символов встречается в названии латвийского интернет-шоу *«This is Хорошо»*.

На данном этапе можно сделать вывод, что использование только графического подхода оформления текстов социальных сетей с малой долей вероятности произведет впечатление на целевого читателя. Рассмотренные примеры наглядно демонстрируют, что чаще всего пользователи при написании интернет-текстов выбирают несколько способов их оформления, например, комбинацию графического и орфографического подходов.

Стандартные правила орфографии при общении пользователей также меняются в сторону упрощения, примером является так называемый язык SMS, т.е. явление, при котором текстовое сообщение изобилует аббревиатурами и сокращениями [Бодулева, Зарипова, 2016, с. 77]: *«Добрый день,соседи!Когда-то обсуждалось,но не могу найти...где ближайший паспортный стол к ЖК Лондон?Номер тлф подскажите пжст?)»*[13]. В этих предложениях используется несколько сокращенных форм слова, а также неполная форма эмотикона *«:-)»*, которая передает положительный настрой.

К языку SMS относятся и *акронимы* – сокращения, появившиеся в результате слияния начальных букв (частей) слов или словосочетаний: *«лс»* – личные сообщения, *«пмсм»* – по моему скромному мнению и пр.: *«...за подбором, подробностями в лс Наталья Попкова или ждем Вас в нашем уютном офисе в г. Всеволожск!»*[14] [López-Rúa, 2007; Кувшинская, 2014]. С. Херринг также отмечает, что наравне с сокращениями используется и другой прием орфографического

[13] https://vk.com/kudrovolife?w=wall-51766355_401437

[14] https://vk.com/mustangtravel?w=wall-438355_46955

оформления текста, а именно – употребление сокращенной формы слова, которая позволяет сэкономить время при написании поста или сообщения и внести тональный оттенок [Herring, 2012]. В русском языке к таким сокращениям относятся фонетически редуцированные слова, например, *«привет»* будет произноситься в разговорной речи как «прет» /ˈprʲɵt/, *«сейчас»* примет форму «щас» /ˈɕ:æs/ и т.д.: *«Строителей 20к2, у вас там кот щас улетит!»*[15].

Необходимо учесть, что социальные сети и интернет в целом не являются единственным источником, который влияет на форму написания слова. L. Squires отмечает, что на орфографический вид слова большое влияние оказывает стиль текста и его тематика [Squires, 2010, p. 463]: *«Галочка, ты щас прочтёшь: ... в мире зафиксировано уже более 150 000 000 случаев заражения коронавирусом...»*[16]. Ключевая идея данного сообщества заключается в неформальном представлении информационных текстов. Использование измененной цитаты в начале новости, в которой пользователь встречает рассмотренную ранее редуцированную форму слова *«щас»*, обусловлено не столько эмоциональным порывом, сколько формой представления новостной сводки.

Итак, пользователи социальных сетей прибегают как к графическим, так и к орфографическим способам оформления интернет-текстов. При их обработке, в частности, при создании модели *Bag-of-Words* для дальнейшего ее представления в векторной форме, подобные способы сложно учитывать. Согласно [Некрасова, Гусев, 2022], стандартный процесс автоматической обработки естественного языка независимо от жанровой принадлежности текста включает такие этапы, как удаление нетекстовых элементов (в том числе эмотиконов) и сведение всех языковых единиц к нижнему регистру. Лингвистические модели чувствительны к искаженным входным данным [Moradi, Samwald, 2021], поэтому, например, при сохранении исходного регистра, который может передавать общее настроение текста, или замене эмотикона на соответствующую эмоцию, переданную словами, потенциально может быть получена некорректная векторная модель текста.

---

[15] https://vk.com/wall-51766355_2864487

[16] https://vk.com/lentach?w=wall-29534144_15341639

1.2 Морфосинтаксические особенности интернет-текстов

Нарушение основных морфологических правил во время интернет-коммуникации проявляется в искаженном употреблении морфологических форм. Например, категория падежа представляет сложность даже для носителей русского языка: *«Ребята, которые пользуются платформой Steam, мб кто-то сталкивался с проблемой, что по прибытию в ПУНК стим перестаёт подсоединяться к серверам и предлагает работать только в автономном режиме?»*[17] Автор поста использует не только акроним *«мб»*, но и нестандартный вариант предложной конструкции с нарушением предложного управления *«по прибытию»* (в сочетании с дательным падежом), хотя в данном случае предлог *«по»* в значение *«после какого-либо момента»* времени употребляется с существительным в предложном падеже.

Морфологические изменения происходят и для *заимствований* – лексем, морфем или синтаксических конструкций, которые были перенесены в результате лингвистических контактов из одного языка в другой [Ярцева, 1998]. Заимствования также характерны для интернет-текстов. В примере *«...а квартиру переоборудует под притон бизнес-тренеров, коучей и прочих демонических отродий...»*[18] выделенное слово заимствовано из английского языка, оно используется для обозначения тренера или наставника. Это слово фонетически адаптировалось к русскому языку [Мамаев, Зайцева, 2019, с. 111], а также приобрело морфологические характеристики имени существительного мужского рода. В некоторых случаях передача заимствованных единиц происходит при помощи фонетических возможностей заимствующего языка [Тюленева, 2016, с. 101]. Например, слово *e-mail* в русском языке приобрело вариант *«мыло»: «Для получения более подробной информации пишите на мыло»*[19].

В исследовании [Крылова, 2016] также отмечается, что морфологическим средством оформления интернет-текстов являются междометия, передающие

[17] https://vk.com/overhearspbsu?w=wall-58219172_160527
[18] https://vk.com/kudrovolife?w=wall-51766355_1856437
[19] https://comp_slang.academic.ru/161/%D0%BC%D1%8B%D0%BB%D0%BE

субъективные состояния пользователей, такие как радость, удивление, гнев и др.: *«А вот вопросик : первому,поздравившему Вас,дорогой Веселый 1го числа с днем рождения...мммм..подарок будет?»*[20]

Для указания на усиление интенсивности значения в текстах социальных сетей используется *редупликация*, которая в «Лингвистическом энциклопедическом словаре» [Ярцева, 1998] определяется как «фономорфологическое явление, состоящее в удвоении начального слога (частичная редупликация) или целого корня (полная редупликация)». Анализ современных примеров из онлайн-бесед указывает на усиление этой тенденции. Так, во время исследования языка социальных сетей в Камеруне C.A. Tabe установила, что в исследовательских данных наблюдались два типа редупликации: во-первых, пользователи прибегали к полному повторению всего слова или его сокращенной формы, во-вторых, они повторяли некоторые буквы в слове для оказания определенного влияния на собеседника:

B: accept greetings too from anastasia and chao

A: ooooooooooooooooooooooooohnooooooooooooooooowhats up with ma girl… [Tabe, 2018, p. 15]

Воняло ужасно,я сначала подумала,что просто разлили что-то вонючее и это припекло от Жары,но нееет...[21]

Отдельно стоит уделить внимание синтаксическим характеристикам интернет-текстов, которые, по мнению лингвиста S. Herring, не так детально описаны в современной литературе [Herring, 2012]. Она утверждает, что одна из характерных синтаксических особенностей – использование высказываний, которые указывают на действие или состояние человека: **молчит*, *плачет** [Herring, 2012]. Подобные высказывания оформляются как синтаксически, как и графически с помощью соответствующих эмотиконов.

Явления синтаксического уровня подвергаются упрощению. Е.В. Холодковская утверждает, что во время интернет-общения характерны

[20] https://vk.com/veselyposelok?w=wall-9544_109156
[21] https://vk.com/kudrovolife?w=wall-51766355_2826460

«языковая краткость, меньшая нормированность, использование эмоционального синтаксиса, упрощение синтаксических структур и влияние синтаксиса разговорной речи» [Холодковская, 2014, с. 82]. Среди приемов, помогающих достичь краткости пользовательского сообщения, выделяют использование простых предложений и неполных предложений, эллипсис и инверсии: *«Красиво: От 5 до 10 ночей в Стамбуле из Москвы 11900-16500 рублей с человека! Вылеты 7 и 8 декабря»*[22]. Отдельно стоит отметить и *парцелляцию* – «экспрессивный стилистический прием, который делит предложение на самостоятельные отрезки, графически выделенные двоеточием, точкой, многоточием и т.п.» [Иноземцева, 2011, с. 114], и помогает привлечь внимание пользователей социальных сетей к важной информационной составляющей: *«Не отказываемся: Слетать на недельку в Турцию из Петербурга за 11000 рублей с человека! Или на две недельки - за 17200 с каждого»*[23].

Использование синтаксических конструкций с «нанизыванием» однородных членов предложения отражает эмоциональную окраску поста [Юйси, 2021]: *«Желаю здоровья, удачи, любви, везения, мира, добра, улыбок, благополучия. Желаю солнца, тепла, мира, веселья, денег, успехов во всех начинаниях, любви, благополучия, исполнения самых заветных желаний, здоровья и вдохновения!»*[24]. В данном примере пользователь комбинирует прием синтаксического параллелизма с использованием большого количества однородных дополнений, что позволяет передать искренность новогоднего пожелания.

Другая отличительная особенность интернет-текстов – нарушение логико-синтаксической целостности предложения, при которой невозможно выстроить связи между частями предложения или установленные связи будут неоднозначно интерпретированы [Алексеева, 2014, с. 9]: *«В нашем подъезде, на первом этаже часто оставляют ненужные вещи, книги, игрушки, бывает обувь и тд. Но то что оставили сегодня….два больших пакета сушеной травы. Моя фантазия вышла из*

[22] https://vk.com/vandrouki_tours?w=wall-121258913_139156
[23] https://vk.com/vandrouki_tours?w=wall-121258913_139167
[24] https://vk.com/wall16825336_999

чата»[25]. Выделенное незаконченное предложение нарушает изначально заданную тема-рематическую связь предложений. У читателя может возникнуть вопрос, какое чувство хотел передать автор именно в этом предложении.

Пунктуационное оформление текстов в социальных сетях имеет свои характерные черты: авторы постов отклоняются от норм постановки знаков препинания в простых и сложных предложениях, опускают тире между подлежащим и именным сказуемым, путают дефис и тире, а также виды кавычек при наборе сообщения, ошибаются при обособлении причастных и деепричастных оборотов [Алексеева, 2014, с. 9]: *«Вопрос к жителям Веселого Поселка, кто городской очередник и стоит на программе "молодежи - доступное жильё", и "развитие долгосрочного жилищного кредитования"...»*[26]. С точки зрения пунктуации пользователь отклонился от следующих пунктуационных норм. Во-первых, была использована запятая между однородными дополнениями. Во-вторых, автор перепутал дефис и тире при оформлении эллипсиса. Наконец, он использовал кавычки-"лапки", которые в русском языке используются как кавычки второго уровня – это случаи, когда необходимо выделить лексические единицы кавычками внутри текстового массива с кавычками.

1.3 Прагмасемантические особенности интернет-текстов

Параллельно с изучением морфосинтаксических параметров постов социальных сетей растет потребность в детализации особенности коммуникации между пользователями. Например, в статье M. Mehmet и его коллег взаимодействие между пользователями изучается на основе расширения существующей теории функциональной лингвистики, которая названа мультимодальной социосемиотической структурой (*Social Semiotic Multimodal framework, SSMM*) [Mehmet, Clarke, Kautz, 2014]. Авторы пришли к выводу, что грани между пользовательскими ролями создателя сообщения и реципиента стираются, т.е. они

[25] https://vk.com/kudrovolife?w=wall-51766355_2985260
[26] https://vk.com/veselyposelok?w=wall-9544_1058334

находятся в одном семиотическом пространстве и могут вести общение без проблем.

Для привлечения большого количества коммуникантов в обсуждение проблемы в конце постов люди могут использовать *хэштег* – «слово …, начинающееся с символа #, служит для пометки сообщения о его принадлежности к какому-либо событию, теме или обсуждению» [Кан, 2017, с. 91], т.е. он дополняет основную семантическую составляющую текста. К. Скотт рассматривает прагматический вклад хэштегов в социальную сеть Twitter[27] [Scott, 2015]. С точки зрения теории релевантности автор констатирует, что хэштеги наводят читателей на определенные идеи и, таким образом, помогают им сформировать наиболее полное представление о тематике твита. Также хэштеги приводят к тому, что пользователи чаще всего стали предпочитать неформальный стиль общения даже без учета иерархии социальных и коммуникативных ролей.

Для усиления коммуникативного намерения говорящего пользователи сопровождают тексты сообщений и постов мемами. *Мем* – это «использующийся в коммуникации знак, имеющий устойчивую форму, которая содержит изменяющийся концепт», он обладает следующими основными характеристиками: эмоциональность, минимализм и актуальность [Смородина, 2019, с. 78-79]. В отличие от стандартных лингвистических особенностей интернет-общения, мемы могут быть поняты определенными категориями пользователей [Щурина, 2012]. Мем, представленный на рисунке 1, могут воспринять те лица, которые:

— владеют русским и английским языками;

— способны расшифровать аббревиатуру SVO – subject, verb, object (подлежащее, сказуемое, дополнение);

— разбираются в типологии порядка слов и способны противопоставить SVO порядки, характерные для других языков: например, VSO характерен для ирландского языка.

[27] Платформа Twitter, на которую приводятся ссылки в данной работе, заблокирована на территории РФ.

Рисунок 1 — Пример лингвистического мема[28]

Решение о включении или исключении изображений из разрабатываемых корпусов основывается на целях и задачах исследования. Во-первых, если основное внимание уделяется машинному обучению на текстах или статистическому анализ, то наличие изображений окажется избыточным как с точки зрения объема хранимых данных, так и использования ансамблей специальных программ. Во-вторых, учет мультимодальности характерен преимущественно для национальных корпусов, что, например, представлено мультимедийным подкорпусом[29] в составе Национального корпуса русского языка или устным компонентом[30] в Британском национальном корпусе или же для профильных корпусов, например, корпуса русского жестового языка [Кагиров и др., 2020]. В составе веб-корпусов представлены тексты, а весь процесс их очистки сведен в документации (см., например, информацию[31] о Генеральном Интернет-корпусе русского языка).

Выводы по первой главе

На данный момент тексты интернет-пространства представляют собой обширное поле для проведения лингвистических исследований. Многочисленные

---

[28] https://vk.com/feed?q=%23%D0%BA%D0%B0%D0%BF%D1%80%D0%B8%D0%B7%D0%BD%D1%8B%D0%B9_%D1%80%D1%83%D1%81%D1%81%D0%BA%D0%B8%D0%B9&section=search&w=wall-116496582_87172

[29] https://ruscorpora.ru/page/manual-murcosearch/

[30] https://cass.lancs.ac.uk/cass-projects/spoken-bnc2014/

[31] http://www.webcorpora.ru/%D0%BE-%D0%BA%D0%BE%D1%80%D0%BF%D1%83%D1%81%D0%B5

работы затрагивают как общие вопросы языка в онлайн-среде, так и отдельные аспекты. Для современного интернет-пользователя язык – это неотъемлемая потребность и необходимость в выражении эмоций при общении. В интернет-текстах пользователь обращается к их аналогам – визуальным, грамматическим и др. Последние инновации в интернет-технологиях позволяют расширить возможности для создания более выразительного текста, что, конечно, представляет интерес для последующего анализа. Тем не менее при работе с текстами в сети Интернет, в том числе при создании корпуса, нужно учитывать, что:

1) ненормативные конструкции, используемые пользователями, могут свидетельствовать не только о нарушениях речевых норм, но также и о богатых возможностях языка;

2) существует уникальный лексический пласт, который может включать в себя новые слова и нестандартные сокращения, затрудняющие коммуникацию или автоматическую обработку текстов;

3) графическое оформление текстов сопровождается внедрением большого количества нестандартных символов или их комбинаций.

С одной стороны, подобное комбинирование лингвистических и экстралингвистических элементов в интернет-текстах помогает достичь максимального эффекта на читателя, однако с точки зрения компьютерной лингвистики, все эти факторы усложняют обработку интернет-текстов, что влечет за собой необходимость постоянного улучшения не только стандартных процедур предобработки текстов, но и морфосинтаксических и семантических анализаторов. На данный момент сочетание автоматизированных и ручных методов является более полезным подходом при обработке веб-корпусов.

Глава 2. Теоретические основания междисциплинарных исследований скрытых сообществ

2.1 Понятие *‘скрытые сообщества’*

Внутренняя структура социальных сетей давно находится в центре внимания специалистов по медиакоммуникации, социологов и лингвистов, при этом основные исследования направлены на описание нечетких или скрытых связей. Подробная характеристика структуры сообществ не только способствует выявлению наиболее важных параметров коммуникации и поведения людей, но и помогает увидеть общую картину сетей. Результатом ряда исследований стало формирование такого важного термина, как *‘скрытые сообщества’*.

В современной терминологии встречается ряд близких терминов, указывающих на использование скрытых сообществ в той или иной дисциплине. Так, академические социальные сети выполняют роль места для общения и взаимодействия ученых, они также помогают установить и укрепить научную репутацию исследователей. Академические социальные сети также предоставляют ценные данные о взаимодействии ученых, между которыми устанавливаются неявные связи *невидимого колледжа*. Идея о *невидимых колледжах* впервые появилась в 1960-х годах благодаря Дж. Прайсу. Он использовал этот термин для описания неформальных взаимоотношений между продуктивными учеными в определенной области, они оказывают влияние на создание новых трудов [Прайс, Бивер, 1976]. Однако в 1960-х годах гипотеза *невидимых колледжей* была опровергнута. Н. Маллинз указал на то, что научная среда организована как распределенная коммуникативная сеть, а не как группы с сильными связями [Маллинз, 1976]. Другие исследователи, такие как Д. Крейн, также обнаружили, что гипотеза не учитывает взаимодействия между активными и обычными участниками научной сети, а также влияние «посторонних» агентов на социальную организацию науки [Крейн, 1976]. Другим близким термином является *‘скрытые сообщества по интересам’* (*hidden communities of interest*). В [Malaterre, Lareau, 2024] они определяются как группы акторов, публикующие близкий с точки зрения

семантики контент в социальных сетях, но обладающие нечеткими социальным связями.

При рассмотрении термина *‘скрытые сообщества’* в настоящей работе необходимо отметить, что хотя S. Fortunato и утверждает, что сформулировать общее определение сообщества достаточно сложно из-за неоднородности используемых систем, датасетов и анализируемых свойств [Fortunato, 2010], тем не менее, эти параметры являются уже частными случаями, эти параметры закладываются отдельным исследователем. В узком смысле под скрытым сообществом понимается *клика* – подграф, все вершины которого смежны друг с другом [Alba, 1973; Fortunato, 2010]. В широком смысле *скрытые сообщества* – это группы пользователей социальных сетей с общими интересами, связи между которыми либо размыты, либо отсутствуют. Скрытые сообщества в веб-пространстве можно сравнить с реальными сообществами с нечеткой организационной структурой: социально опасные группы наркозависимых, секретные организации и пр. В отличие от сплоченных сообществ наподобие семейств, коллег или лучших друзей, связи между членами вышеупомянутых скрытых сообществ в реальной жизни являются неочевидными [He et al., 2015; He et al., 2018; Mamaev, Mitrofanova, 2020a]. Скрытые сообщества образуют своего рода независимые слои внутри целостного графа, из этого следует, что понимание структур скрытых группировок, а также их организации имеет решающее значение для более комплексного описания исследуемой системы.

Формальная структура скрытых сообществ в онлайн-пространстве чаще всего представляется в виде семантической сети предметной области, для которой разрабатывается уникальный алгоритм обнаружения: скрытые группы выделяются на основе расстояний между вершинами (*feature distance*), внутренней плотности (*internal density*) и т.д. [Coscia, Giannotti, Pedreschi, 2011]. Однако большинство алгоритмов имеют общую процедуру построения, которая описывается следующим образом.

1. С некоторого онлайн-пространства происходит парсинг формальных пользовательских идентификаторов, которые будут являться узлами. В них также может содержаться метаинформация: пол, возраст, место рождения и др.

2. Детализируются условия незнакомства пользователей, проверяется их истинность.

3. На основании общности ряда лингвистических и математических параметров строятся ребра, соединяющие пользователей. К лингвистическим параметрам объединения можно отнести общность тематических блоков, которые формируются при публикации текстов постов на личных страницах.

Итогом этих операций будет *семантическая сеть* (рисунок 8). Семантические сети можно охарактеризовать как формальное представление знаний, приобретающее вид ориентированного графа, в котором объекты или ситуации представляются вершинами, а отношения между ними – дугами (ребрами). Семантическая связь отражает отношения между понятиями. В лексической системе языка отношениям соответствуют единицы любого вида, включая предикаторы «меньше», «если, то» и др. [Гущин, 2007]. Для описания связей в социальных сетях используются отношения вида «быть другом», «быть членом тайного сообщества» и др. На рисунке 2 оранжевым цветом обозначены вершины – центры явных сообществ, в которых пользователи знают друг друга или же, по крайней мере, находятся в категории «Друзья» в какой-либо социальной сети. Цветные ребра и подузлы уже формируют сообщество с неявными связями.

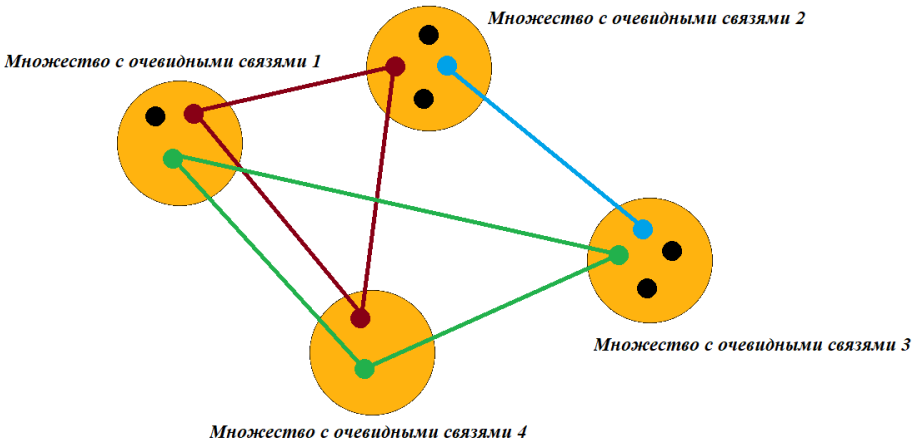

Рисунок 2 — Графическое представление модели скрытых сообществ

Построение подобных моделей активно применяется в современных технических и социогуманитарных дисциплинах. Например, в области социологии объектом исследования П.А. Мейлахса и его коллег попала такая группа пользователей социальных сетей, как СПИД-отрицатели [Мейлахс, Рыков, 2015; Рыков, Кольцова, Мейлахс, 2016]. Авторы с помощью метода «обоснованной теории» (*Grounded theory*) закодировали посты и провели их детальный анализ, акцентируя внимание на используемых дискурсивных стратегиях, которые СПИД-диссиденты применяют для того, чтобы убедить сомневающихся в этих идеях личностей. Эти стратегии можно свести к следующим: «научные» аргументы, описание личного опыта, идеологически направленные аргументы и т. д. В результате пользователи скрытых сообществ СПИД-диссидентов после прочтения постов в тематически ориентированных группах подвергались сильному влиянию. Они вступали в эти группы, а термин *‘скрытый’* менялся на *‘явный’*.

О.С. Смирнова рассматривает проблему идентификации пользователей детского и подросткового возрастов, которые склонны к суицидам. Материалом анализа послужили детские и подростковые посты, опубликованные в открытом доступе на сайте ВКонтакте. Выделенные признаки суицидального поведения классифицируются и оцениваются. Автоматизированное выявление подобных сообществ позволяет оградить подростков от негативного суицидального контента [Смирнова, 2017].

Еще одна решаемая задача социологического исследования скрытых сообществ – задача выявления опасных сообществ, участники которых имеют пристрастие к запрещенным веществам. Например, С.А. Митягин описывает подробную процедуру выявления пользователей, склонных к наркомании, а также характеризирует выявленные тематики постов [Митягин, Якушев, Бухановский, 2012]. Авторская методика состоит из нескольких этапов.

1. Сперва отбираются посты, посвященные потреблению наркотических веществ, и составить список пользователей, которые являются авторами данных текстов.

2. После перехода по ссылке на страницы пользователей авторы автоматически извлекали мнения, а потом проводили оценку постов. Поиск прекращается, если алгоритм не обнаружил необходимые мнения.

Таким образом, данные социальных сетей помогают смоделировать процесс распространения наркокультуры как в онлайн-обществе, так и в реальном мире. Наравне с подобным сообществом ученые также рассматривают скрытые сообщества проституток, гомосексуалистов, бездомных и др. [Печенкин, Зайонц, 2011; Wejnert, 2009].

Стоит отметить, что группировки суицидентов и наркоманов до внедрения интернет-технологий на территории Российской Федерации могли формироваться только в реальной жизни, сегодня же их распространение повсеместно. В отличие от них спамеры как социальная группа имеет отправную точку в онлайн-среде, она существует именно там. В работе [Bindu, Mishra, Thilagam, 2018] акцентируется внимание на том, что спамеры, как правило, формируют сообщество спам-аккаунтов и используют их для распространения ненужных электронных посланий среди большого числа рядовых пользователей, поэтому скрытые сообщества спамеров как онлайн-явления представляют наибольший интерес для социологов. В исследовании предлагается алгоритм обучения без учителя под названием SpamCom для обнаружения спамерских сообществ в социальной сети Twitter, которая интерпретируется как многоуровневая платформа. Сеть пользователей представлена авторами в виде мультипараметрического графа, на основе анализа которого выявляются стандартные стратегии, используемые спамерами. Авторы утверждают, что комбинация различных параметров (характеристика URL-адресов учетных записей пользователей, семантическая близость контента пользователей и пр.) являются прочной основой для моделирования скрытых сообществ спамеров.

Еще одним потенциально опасным сообществом является группировка приверженцев радикальным взглядам. В исследовании [Agarwal, Sureka, 2015] под вышеупомянутыми приверженцами понимаются пользователи, для которых характерно распространение ненависти и экстремизма, поэтому перед учеными была поставлена задача на основе реального корпуса блогов Tumblr построить граф

и выявить подграфы скрытых сообществ недоброжелателей. Авторами был предложен собственный краулер, выполняющий сразу несколько задач: поиск блогера, вычисление его сходства с образцовыми документами, фильтрация блогеров, пропагандирующих ненависть, а также навигация по ссылкам на других блогеров (на основе перепостов и лайков). Экспериментальные данные были вручную проаннотированы 30 студентами-выпускниками, на основе которых уже и принималось решение о включении пользователя в скрытое сообщество. F-мера, равная 80%, доказывает работоспособность алгоритма, а ключевой особенностью потенциального вхождения пользователя в экстремистское сообщество является именно перепост.

Задача моделирования скрытых сообществ рассматривается также и в области лингвистики. Статья [Хорошевский, Ефименко, 2017] посвящена выделению скрытых сообществ в научных кругах, материал исследования – электронные версии статей конференции «Диалог», опубликованные в период с 2000 г. по 2009 г. При изучении текстовых коллекций научных трудов и способов цитирования можно выяснить, как организованы скрытые исследовательские сообщества – группы людей, использующие в своих экспериментах одни и те же или схожие наборы данных. В результате авторы предсказали укрупненные тематические блоки, которые будут популярны среди исследователей-разработчиков в обозримом будущем (например, синтез и генерация устной речи, обработка текстов метода искусственного интеллекта и пр.). Анализ научных публикаций описан также в [Lafia et al., 2022]: в этой статье анализируется структура научного сообщества на основе работ по социальным дисциплинам, которые были предоставлены Межуниверситетским консорциумом политических и социальных исследований (*the Interuniversity Consortium for Political and Social Research, ICPSR*). В ходе работы было выделено 41 скрытое сообщество, а детальный анализ структуры групп позволяет утверждать, что размер самого сообщества прямо пропорционален уровню используемости одного и того же набора данных: чем больше узлов графа в группе, тем более важным является

используемый учеными набор данных. Таким образом, исследование позволяет переосмыслить подход к отбору данных при проведении экспериментов.

2.2 Алгоритмы определения скрытых сообществ

Скрытые сообщества, как уже было подчеркнуто в предыдущем разделе, играют важную роль при анализе внутренней структуры социальных сетей. В зависимости от цели исследования учеными различных дисциплин одни и те же алгоритмы могут существенно модифицироваться, в результате чего их сложно сравнивать. Однако, как указано в [Mamaev, Mitrofanova, 2020b], существующие алгоритмы можно разделить по основанию, на котором развиваются те или иные процедуры: подходы на основе графово-математического анализа, подходы с привлечением кластерных алгоритмов и гибридные подходы (например, сочетание графовых подходов с привлечением лингвостатистической информации). Поэтому в данном разделе представлено описание алгоритмов в соответствии с указанной классификацией.

Первая классификационная группа – графовая. Под *графом* в широком смысле этого термина обычно понимают совокупность взаимосвязанных узлов, более строгое определение сформулировано следующим образом: графом G называется пара множеств $\langle V, E \rangle$, где V – это множество вершин, а E – множество ребер [Баранский, Расин, 2008, с. 4]. Развитие теории графов начинается с 1736 года, когда Л. Эйлер анализировал задачу Кёнигсбергских мостов [Euler, 1741]. В дальнейшем к научному исследованию графов обратились в XIX веке, когда в 1847 году немецким физиком Густавом Кирхгофом была написана статья по поводу решения систем уравнений для определения силы тока в электрических цепях, после чего Кирхгофа стали считать родоначальником теории деревьев [Харари, 2003; Kirchhoff, 1847]. С тех пор графы рассматривались как объект с определенными математическими свойствами [Bollobas, 1998]. В XX столетии теория графов стала одним из популярных методов формализации данных различных предметных областей. Например, социальные сети и интернет-сообщества в реальных условиях формально представляются в виде графа, однако

быстрое разрастание такого графа становится проблемой сетевого анализа. Таким образом, возникает острая необходимость в фундаментальном переосмыслении графовых представлений данных с помощью современных инструментов [Pastor-Satorras, Vespignani, 2004; Newman, 2003], в том числе и в лингвистических дисциплинах. В этой связи для анализа данных возникает большое количество методов с различными подходами: метод кратчайшего незамкнутого пути, метод поиска ядра графа, метод поиска минимального покрывающего дерева и др. [Малафеев, Щеникова, Скворцова, 2021]. Например, метод кратчайшего незамкнутого пути заключается в разбиении исходного графа на некоторое число кластеров N, которое пользователь вводит заранее. Сначала производится поиск кратчайшего незамкнутого пути, а затем происходит удаление $(N-1)$ ребра с максимальным весом. Данный путь будет наикратчайшим и незамкнутым при условии, если общее значение весов ребер окажется минимальным и не будет существовать циклов. Если обратиться к методу поиска ядра графа, то с помощью алгоритма Магу производится поиск подмножеств множества вершин исходного графа, узлы которого являются как внешне, так и внутренне устойчивыми. Внешне устойчивым считается такое множество вершин, в котором либо любая вершина входит в данное множество, либо вершина не входит в анализируемое множество, но из этой вершины в данное множество ведет некоторая дуга. Внутренне устойчивым считается такое множество, в котором каждый элемент не является смежным по отношению к любому другому элементу.

Графы социальных сетей как вид социальных графов отличаются стремительным увеличением количества узлов и ребер, что может привести к неоднородности отображения структуры [Задорожный, Юдин, 2017]. В исследовании M. Wang и коллег [Wang et al., 2023] предлагается следующее решение: чтобы справиться с быстрым ростом размера графа, авторы пытаются обнаружить скрытые сообщества с помощью метода, который совершает определенное количество итераций, чтобы выбрать новый подграф из исходной сети для дальнейшего анализа. Сначала из одного узла на основе

модифицированного локального спектрального метода происходит расширение окружения (т.е. производится поиск возможных узлов локального сообщества), результатом данного этапа является обнаружение исходного доминирующего локального сообщества. На втором этапе узлы (члены) данного сообщества и их связи с другими узлами удаляются. Итогом является выделение соседних сообществ в оставшемся подграфе, в том числе некоторые «неполноценные сообщества» (англ. *broken communities*), которые содержат только часть членов исходной сети. Локальное сообщество в совокупности с соседними сообществами образуют доминирующий слой. Уменьшая веса ребер внутри каждого из этих сообществ, авторы ослабляют структуру изучаемого слоя, таким образом, выявляются скрытые слои. Совершив n-ное количество итераций, можно обнаружить все скрытые сообщества, содержащие начальный узел.

Авторы показывают, что предложенный метод теоретически может использоваться при работе со сложными по структуре сообществами: например, если в исследуемом сообществе присутствуют «разорванные» связи. Данный метод может значительно превзойти самые современные базовые алгоритмы, которые были разработаны либо для глобального обнаружения скрытых сообществ, либо для обнаружения локальных сообществ.

В исследовании [Gmati et al., 2018] был описан процесс разработки метода выявления скрытых сообществ *Fast-Bi Community Detection* (*FBCD*). В основе этого метода лежит двудольный граф, который состоял из двух наборов узлов и соединяющих эти узлы ребер. Основная идея заключается в том, чтобы изучить множество максимального соответствия в двудольном графе с уменьшением сложности самого алгоритма. Практическим материалом исследования послужили корпуса различных социальных сетей, в том числе и *Facebook*[32]. В результате были выявление такие скрытые сообщества, как American Revolution, Southern Women и некоторые другие.

[32] Организация Meta, а также ее продукты Instagram и Facebook, на которые приводятся ссылки в настоящей работе, признаны экстремистскими на территории РФ.

Графовые методы применялись A. Hopkins при выявлении скрытых связей в сети «Аль-Каида»[33] [Hopkins, 2010] (рис. 3). В качестве базовых наборов данных использовались различные версии правительственных публикаций, статей и книг по данной тематике. Подсчитанные параметры глобальной эффективности и средних расстояний между узлами позволили автору утверждать, что данную сеть можно описать как «темную сеть» (*dark network*).

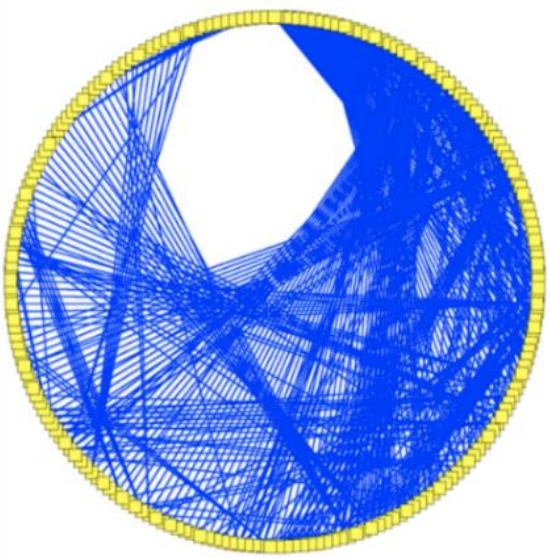

Рисунок 3 — Итоговая сеть «Аль-Каиды»[34] из работы [Hopkins, 2010]

В [He et al., 2018, p. 92] описана разработка алгоритма выявления скрытых сообществ *HIdden COmmunity DEtection*, получившего условное название *HICODE*. Авторы разбивают данную процедуру на два этапа, первый из которых нацелен на определение количества слоев сообществ, обозначающий уровень скрытости исследуемой группы. При обнаружении первого слоя можно сделать вывод, что данная группа обладает максимальной степенью связности, при этом каждый следующий слой меньшей степенью связности. Понятие *связности* определяется следующим образом. Пусть $G_i = G(V_i)$ – подграф, порожденный множеством вершин V_i, $(1 \leq i \leq k)$. Графы $G_1, G_2, \dots, G_k$ называются компонентами связности графа G. Граф, обладающий ровно одной компонентной связности, называется связным [Баранский, Расин, 2008, с. 9-10].

[33] Данная организация, на которую приводятся ссылки в работе, признана экстремистской на территории РФ.

Следующий этап «улучшает» качество выявленных слоев, он называется уточняющим этапом. Авторы утверждают, что к нему необходимо прибегнуть, так как более сильная структура сообщества может привести к искажению сведений о структуре сообществ с более слабой структурой [He et al., 2018, p. 96]. Ранее схожая мысль была высказана в работе [Young et al., 2012]. При построении графов крупных размеров значительная часть информации не вносится в сеть. Данный недостаток возникает из-за того, что сообщества с большой плотностью могут «перекрыть» более разреженные группировки. Авторы предложили общий каскадный подход к обнаружению скрытых сообществ, позволяющий снизить процент неупомянутых сообществ.

Алгоритм *HICODE* показал свою эффективность при обнаружении скрытых сообществ Reddit. В исследовании [Salz, Benavides, Li, 2019] авторы предложили усовершенствовать данную процедуру, добавив в нее алгоритмы предсказывания границ группировок и весов ребер. Было установлено, что из-за больших размеров итогового графа и активным взаимодействием между пользователями сайта границы сообществ приобретают размытые очертания.

В науке о сетях наравне со скрытыми сообществами могут изучаться и отдельные сообщества, которые при первичном построении модели сообществ не были выявлены исследователем. В англоязычной традиции данная проблема получила название *deep community detection* – выявление глубинного сообщества. Под *глубинным сообществом*, согласно определению P.Y. Chen, понимают связанный компонент, который выявляется только после удаления ребер и узлов из остальной части исходного графа. В работе [Chen, Hero, 2015] процедура обнаружения глубинных сообществ – это многоэтапный процесс удаления узлов, в результате которого выявляется новая мера центральности графа. Данную меру в статье называют локальной центральностью вектора Фидлера (*local Fiedler vector centrality, LFVC*), одна из ее особенностей – чувствительность алгебраической связности к удалению ребер/узлов. Жадная стратегия LFVC при увеличении исходного графа может извлекать глубинные сообщества с вероятностью, близкой

к 100%. Алгоритм апробируется на ряде набор данных: *Dolphin social network*[35], *Zachary's karate club*[36] и др. По сравнению с обычными методами обнаружения сообществ авторы показывают, что их алгоритм лучше выявляет важные сообщества и их ключевых участников.

Вторая классификационная группа связана с использованием кластерного анализа, т.е. упорядочивания любых фактов и явлений определенных объектов в сравнительно однородные группы по признаку их схожести с дальнейшей интерпретацией [Кутыркин, Сёмин, 2009, с. 5]. Данный подход является одним из основополагающих в современных экспериментах по социологии. Например, в одной из недавних работ A. Wong продемонстрировала, как из неоднородного массива данных о студентах извлечь информацию о социально опасных группировках, в которые входят лица, склонные к суицидальному поведению [Wong et al., 2022].

Если обращаться к задачам обработки естественного языка, то первостепенным объектом кластеризации становится текстовая информация, из которой 80%, согласно опросу, проведенному IDC, имеет неструктурированный вид в веб-пространстве [Chakraborty, Krishna, 2014].

В педагогике развернутые отзывы как вид неструктурированных данных об академической успеваемости студентов играют основополагающую роль в совершенствовании методов преподавания различных дисциплин, что подтверждается рядом исследований [Алисов, 2015; Сапегина, 2020]. Стандартные методы анализа и оценки (анкетирование или фронтальный опрос по окончании обучения) имеют ряд недостатков: например, часть неявной информации может быть упущена. A.M. Abaidullah предлагает использовать кластерный анализ на основе алгоритма *k*-средних для выявления скрытой информации. Для сбора текстовых ответов студентов был подготовлен опросник из 28 вопросов (готовился ли преподаватель к занятиям, какая была у преподавателя система оценивания работ студентов и пр.). По результатам автоматического анализа было произведено

[35] https://networks.skewed.de/net/dolphins
[36] http://vlado.fmf.uni-lj.si/pub/networks/data/Ucinet/UciData.htm#zachary

описание данных. Для первого вопроса (*The semester course content, teaching method and evaluation system were provided at the start*) было выявлено три кластера. У первого кластера центроид (центр тяжести) равен 3.6927, а сам кластер покрывает 63% опрошенных студентов. Этот факт указывает на то, что в начале семестра большинство преподавателей действительно предоставили содержание курса, методику обучения и систему оценивания. Однако 28% студентов во втором кластере указывают на мнение, противоположное тому, что выражается в первом кластере, и оно имеет определенный вес при формировании стратегий преподавания в будущем. Данное мнение, возможно, было бы упущено при ручном анализе анкет студентов [Abaidullah, Ahmed, Ali, 2015].

В работе [Sánchez-Rebollo, 2019] проводится анализ Twitter-постов с целью обнаружения потенциальных лидеров террористических организаций и их последователей. Для проведения процедур нечеткой кластеризации использовался сбор нескольких параметров: содержание постов, уровень активности пользователей и их влияние на других лиц. Автор подчеркивает, что подобным образом сотрудники миграционных служб смогут извлечь дополнительную информацию о людях, с которыми они ведут переговоры в данный момент, а представители страховых компаний смогут определять нежелательных клиентов.

Нечеткая кластеризация как эффективный метод выделения сообществ постулируется и N. Binesh. В процедурах нечеткой кластеризации каждому узлу приписывается процентный показатель его вхождения в данный кластер. В исследовании [Binesh, Rezghi, 2018] предлагается проводить кластеризацию на основе неотрицательного матричного разложения (*non-negative matrix factorization, NMF*). Для заданного параметра c NMF оценивает матрицу данных $X \epsilon \mathbb{R}^{m \times n}$ как произведение двух неотрицательных матриц $V \epsilon \mathbb{R}^{m \times c}$ и $H \epsilon \mathbb{R}^{c \times n}$, опираясь на следующую задачу минимизации (1):

$$min_{V,H \geq 0} \left|\left|X - VH\right|\right|, \quad (1)$$

где $\left|\left|X - VH\right|\right|$ – Евклидова форма матрицы. В задачах кластеризации параметр $c \ll min(m, n)$ принимается как итоговое количество кластеров-сообществ.

Стоит отметить, что для оценки достоверности алгоритмов нечеткой кластеризации введены два новых критерия – *Supervised Fuzzy Evaluation Criterion (SFEC)* и *Unsupervised Fuzzy Evaluation Criterion (UFEC)*, которые строятся на основе структуры соседства узлов и могут оценивать вхождение элементов в сообщество. Экспериментальные результаты на основе корпусов *YouTube* и *Facebook* показали эффективность NFM-кластеризации и достоверность предложенных критериев оценки.

С 2020 г. исследовательское сообщество сосредоточило внимание на текстах, связанных с эпидемией COVID-19. L. Chaudhary собрала корпус текстов о COVID-19 с января по август 2020 г. с официального сайта Университета Джонса Хопкинса. Для выявления скрытых сообществ использовалась комбинация таких методов машинного обучения, как метод главных компонент для уменьшения размерности данных и *k*-средних для выделения кластеров, отображающих скрытую структуру выделенных сообществ стран. Страны и регионы, которые по результатам анализа вошли в одно и то же сообщество, могут оказывать друг другу аналогичную помощь в принятии превентивных мер, чтобы избежать наихудших сценариев развития COVID-19. Сами же скрытые сообщества и данные о них могут пригодится в сфере здравоохранения при формировании дальнейших политических решений [Chaudhary, Singh, 2021].

Необходимо подчеркнуть, что кластерный подход может уступать по эффективности модифицированным графовым подходам. В статье [Jia et al., 2019] авторы выдвигают положение о том, что стандартные алгоритмы кластеризации (например, смешанные гауссовские модели, *k*-средних и т.д.) неприменимы для работы с пересекающимися группами, так как в действительности они обладают тесными связями. Степень устойчивости связей, которые были получены автоматически, гораздо меньше, в связи с чем была представлена процедура *Community Detection with Generative Adversarial Nets* (*CommunityGAN*), чью основу составляют генеративно-состязательные сети. Данные сети основаны на двух нейронных сетях, первая из которых генерирует образцы (вероятные сообщества), вторая выбирает истинные сообщества. Таким образом, уменьшается количество

«шумов» в итоговых моделях, а исследователи получают более четкую информацию о пересекающихся сообществах. Каждой вершине сообщества назначается коэффициент связи с другими сообществами. Результаты экспериментов, проведенные на пяти наборах данных (на основе таких социальных сетей, как YouTube, Amazon и пр.), позволили утверждать, что CommunityGAN может использоваться для решения дальнейших практических задач.

Последний тип алгоритмов – гибридные, сочетающие в себе как уже ранее описанные методы, так и «дополнительные». В [Iqbal et al., 2019, p. 22740] описан эксперимент с применением подобного алгоритма. В фокусе их внимания оказалось скрытое сообщество киберпреступников: эта маргинальная группа благодаря стремительному развитию информационных технологий имеет возможность осуществлять неконтролируемую незаконную деятельность, в том числе кибербуллинг, кибермошенничество и сбыт наркотических веществ. Авторы разработали платформу интеллектуального анализа криминальной информации на основе лингвистического тезауруса WordNet для выявления и извлечения важной для судебной экспертизы информации, хранящейся в чатах. После обработки чата подозреваемого происходит идентификация скрытых подграфов (сообществ) и тем в разговоре каждом подграфе. Два узла связаны, если для ключевых слов каждого пользователя найден общий гипероним в WordNet. В ходе экспериментальной оценки было установлено, что разработанный подход позволяет эффективно извлекать скрытые сообществ и тематические компоненты. Точность выявления группировок и их тематической составляющей в наборе данных по сравнению с другими современными алгоритмами выявления скрытых сообществ повысилась на $10 - 20\%$. Выявленные с помощью WordNet тематики можно сопоставить с существующими базами данных о преступлениях, чтобы выяснить, занимается ли подозреваемый неправомерными действиями.

В исследовании [Zhou et al., 2017] для выявления скрытых сообществ в корпусе *Australian Federal Election 2010* применяется алгоритм тематического моделирования *Latent Dirichlet Allocation* (*LDA*). Для создания тематической модели нужно найти три переменных: оптимальное число тем в корпусе текстов,

распределение $p(t|d)$ для всех текстовых документов, а также распределение $p(w|t)$ для всех тем [Blei, Ng, Jordan, 2003]. Используется формула (2):

$$p(w|d) = \sum p(t|d)p(w|t), \quad (2)$$

где $p(w|d)$ – вероятность появления некоторой лексической единицы w в некотором текстовом документе d, $p(w|t)$ – неизвестная вероятность появления лексической единицы w в некоторой теме t, $p(t|d)$ – неизвестная вероятность появления темы t в документе d.

Авторы отмечают, что объем корпуса мал (около 57000 твитов длиной до 140 символов каждый), а более детальные данные о сообществах для LDA могут быть получены на корпусах бóльших объемов. К тому же в статье не приводится общий граф скрытых сообществ, что немного затрудняет итоговое восприятие результатов. Наконец, для текстов небольшой длины, как отмечается в статье [Chen, Kao, 2015], алгоритмы наподобие LDA или *probabilistic Latent Semantic Analysis* (*pLSA*) часто выводят слова-тематизаторы широкой семантики, поэтому для подобных случаев целесообразно применять модели типа *Biterm Topic Modeling* (*BTM*), которые специально создавались для работы с короткими текстами.

Для исследования сообществ интернет-пользователей используются и нейросетевые подходы. В [Оболенский и др., 2021] авторы обсуждают проблему классификации сообществ в сети ВКонтакте. Они исследуют использование нейронных сетей для классификации групп пользователей с точки зрения степени радикальности. На основе нейронной сети с рекуррентной долговременной памятью (*Long short-term memory, LSTM*) была создана модель, которая обучается на тестовых данных и оценивается с помощью различных метрик – точность, потери и F-мера. На втором этапе была создана модель сверточной нейронной сети, которую сравнили с моделью, разработанной на первом этапе. В результате экспериментов оказалось, что модель на основе сверточных нейронных сетей показывает более высокие значения метрик, чем LSTM-модель. Утверждается, что модель сверточной нейронной сети можно использовать для поиска радикальных сообществ в различных сегментах социальных сетей.

Нейросетевые методы поиска также достигают высоких значений точности определения сообществ по сравнению со стандартными методами машинного обучения. В исследовании [Минаев, 2022] проводится эксперимент по выявлению сообществ экстремисткой направленности в трех корпусах текстов: антисемитизм (268 тыс. словоформ), реабилитация нацизма (284 тыс. словоформ) и радикальный ислам (310 тыс. словоформ). Применяя метод опорных векторов, рекуррентные нейронные сети и модель *BERT (Bidirectional Encoder Representations from Transformers)* из семейства Transformers, автор установил, что последняя модель превосходит оставшиеся с точки зрения точности распознавания скрытых экстремистских сообществ приблизительно на 10%. Подобные результаты должны учитываться при определении информационных воздействий на молодое поколение, особенно в случаях, когда они могут склонять окружающих к суицидальному поведению или вовлекать в опасные сообщества.

В дальнейших исследованиях также применялись комбинации графовых подходов и тематических моделей. В [Mamaev, Mitrofanova, 2020a] создается автор-тематическая модель русскоязычных постов VK, для каждой темы из внешних и внутренних источников выводятся тематические метки, которые в дальнейшем считались основой для создания графа скрытых сообществ. Было установлено, что параметр принадлежности тем к авторам вводится вручную, а не автоматически, как, например, в реализации алгоритма *Author-Topic Modeling* (*ATM*) в библиотеке *gensim*[37]. Предложенный в работе подход можно обозначить как «дискретная автор-тематическая модель». В данном случае «дискретная» обозначает, что идентификация авторства текстов относительно тем происходит до построения общей автор-тематической модели корпуса, т.е. общая автор-тематическая модель складывается из отдельных автономных микромоделей. Так как эксперимент проводился в первую волну коронавируса, то ведущей темой на тот момент стала тема здоровья.

[37] https://radimrehurek.com/gensim/

Выводы по второй главе

Сегодня одним из актуальных направлений научных исследований является импорт больших данных из реальных сетей и моделирование их взаимодействия в виде некоторой структуры, которая может обладать как явными связями, так и неявными. Объекты, объединенные вторым типом связи, можно охарактеризовать как скрытые сообщества – это группы пользователей, которые несознательно находятся в определенной социально-коммуникативной сети благодаря наличию общих элементов, в том числе и лингвистических. Члены скрытого сообщества могут иметь схожие интересы и мнения, манеры поведения, в то время как они могут не осознавать своей принадлежности к какой-либо группе или не знать о существовании друг друга. Скрытые сообщества возникают в различных онлайн-средах, таких как социальные сети, форумы, блоги и другие платформы.

Важно отметить, что для выделения скрытых сообществ используются методы, основанные на переходе от анализа всей исходной структуры к отдельным подструктурным элементам, – графовые алгоритмы, кластерные алгоритмы и комбинированные алгоритмы с привлечением дополнительной информации. Выбор пути реализации алгоритма зависит не только от конкретной исследовательской задачи, но и от других факторов: распределение социологических параметров в наборе данных пользователей (пол, возраст, образование…), язык-доминанта, используемый для создания постов, и др. Внедрение комбинированных алгоритмов представляет собой инновационный метод оценки правильности выделения скрытых сетевых сообществ. В настоящем исследовании будет представлен метод, объединяющий алгоритмические подходы к выделению сообществ и семантические анализаторы русскоязычных текстов.

Глава 3. Разработка процедуры лингвистического профилирования скрытых сообществ

3.1 Критерии и процедуры построения исследовательского корпуса текстов скрытых сообществ

В данном исследовании используется термин *‘корпус социальных сетей’* – коллекция русскоязычных текстов из открытого сегмента социальных сетей, которая подвергается автоматической лингвистической разметке. Для его составления нужно соблюдать идею *сбалансированности (репрезентативности)* – «достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов, то есть способность отражать все свойства языка или подъязыка» [Захаров, Богданова, 2020, с. 22]. Так как корпус будет создан для целей лингвистического анализа скрытых сообществ, используются следующие критерии отбора материала.

1. Так как фиксация голосовых сообщений в постах (а не личных сообщениях) социальных сетей не получила до сих пор широкого распространения, то планируемый корпус будет полностью *письменным*.

2. Несмотря на то, что в дальнейших разделах проводится лингвистическое профилирование не отдельных пользователей социальных сетей, а целых групп пользователей, объединенных общим тематическим компонентом, необходимо сбалансировать корпус *по полу пользователей*, т.е. соотношение мужчин и женщин должно находиться приблизительно в равных пропорциях.

3. Выборка данных должна быть организована *на основе одной социальной сети*, поскольку совокупность текстов единого онлайн-ресурса можно рассматривать как отдельную языковую модель – подкорпус веб-корпуса со своим набором лингвистических параметров. При попытке комбинирования нескольких социальных сетей может возникнуть ряд методологических вопросов: например, принадлежат ли две страницы разных социальных сетей одному реальному лицу при условии, что на них оставлены разнородные непересекающиеся данные?

4. Посты пользователей должны быть *неустаревшими на момент проведения экспериментов и интерпретации результатов*. В исследовании

[Mamaev, Mamaeva, Axenova, 2022] отмечено, что внешняя ситуация в мире напрямую влияет на реакцию пользователей и публикацию ответных постов в социальных сетях, что влияет на тематические сдвиги корпуса. При анализе корпуса русскоязычных постов за 2018-2022 гг. было показано, что количественные изменения в употреблении тематик произошли на рубеже 2019-2020 годов в связи с началом пандемии коронавируса и последующих за ней событий. В результате для экспериментального корпуса были выбраны посты, опубликованные не ранее 01.01.2020.

5. Сама выборка данных будет разделена по пользовательским подкорпусам в соответствии с авторством для упрощения процесса обработки постов. Так как в главе 2 введено понятие скрытых сообществ, то необходимо, чтобы *число общих друзей между двумя потенциальными членами скрытого сообщества было строго равно нулю*, иначе возникнет противоречие с рассматриваемым понятием.

Весь процесс практического исследования можно представить в виде упрощенной блок-схемы, которая будет подробно прокомментирована в дальнейших разделах. Такое визуальное представление поможет сориентироваться в ходе изложения мыслей и понять последовательность действий (рис. 4).

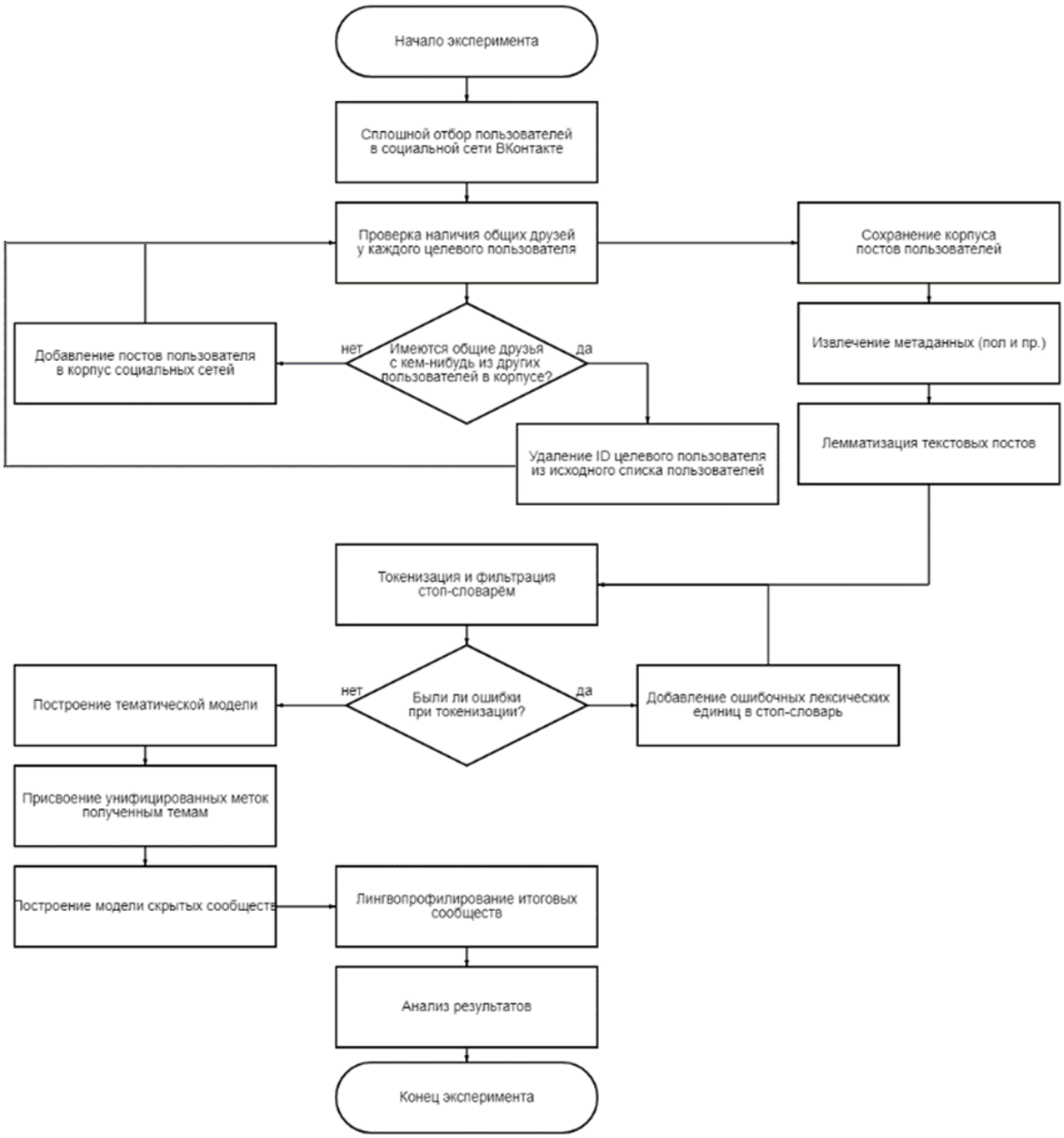

Рисунок 4 — Блок-схема эксперимента

В качестве основной платформы для сбора практического материала был выбран русскоязычный сегмент социальной сети ВКонтакте. Во-первых, согласно исследованию компании Brand Analytics[38], данный ресурс является лидирующим по количеству сообщений и по количеству активных авторов постов на осень 2023 года. Во-вторых, ВКонтакте, в отличие от зарубежных порталов, не требует предустановки специальных программ для аутентификации. Наконец, русскоязычным сообществом программистов разработано большое количество

[38] https://brandanalytics.ru/blog/social-media-russia-autumn-2023/

библиотек для языка программирования Python, которые облегчают процесс выгрузки лингвистической информации и метаданных, в частности, библиотека *vk_api*[39].

Изначально методом сплошной выборки было сохранено более 7000 ID пользователей. Следующим этапом стало создание парсера, в основе которого лежали такие библиотеки как *bs4*[40], *requests*[41] и *vk_api*. На вход подавался список пользователей, для каждого отдельного пользователя в начале списка проводилась проверка с концом списка: являются ли целевой пользователь и пользователь из конца списка явными друзьями в социальной сети. Если они являлись друзьями, то целевой пользователь удалялся из списка. Итоговое количество пользователей – 704.

Для вышеуказанного количества пользователей извлечены посты с опорой на дополнительные требования. Некоторые онлайн-источники[42] приводят показатели минимального количества длины постов для различных социальных сетей. Так, например, для Twitter предлагают использовать 120-130 символов в публикуемых текстах, а для Snapchat желательно не превышать стандартное значение в 80 символов. Было принято решение, что практическим материалом для экспериментов станут тексты социальной сети ВКонтакте длиной более 200 символов, так как для них возможно использование стандартных процедур тематического моделирования и их модификаций.

Наконец, была извлечена доступная метаинформация: имя и фамилия, пол, год рождения, образование, интересы. Оказалось, что по признаку пола данный корпус уже сбалансирован, среди 704 пользователей было 347 женщин и 357 мужчин, при этом 258 пользователей указали свои интересы на странице профиля, что составляет 36.6% от общего числа пользователей. Необходимо отметить, что сохраненные данные не всегда смогут точно и однозначно описывать предпочтения пользователя. На момент создания корпуса у пользователя с ID 241545

[39] https://vk-api.readthedocs.io/en/latest/
[40] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[41] https://requests.readthedocs.io/en/latest/
[42] https://blog.hubspot.com/marketing/character-count-guide

указывалось, что интересы «*у каждого свои*», а пользователь с ID 41961 заявляет, что его интересы «*часто меняются*». Данная информация не дает однозначного ответа на вопрос о пользовательских интересах, поэтому она может быть приравнена к значению *None*, которое присваивается тем пользователям, которые не заполнили информацию о своих интересах.

Тем не менее, для создания модели скрытых сообществ получен обширный материал, состоящий из пользовательских постов, относящихся к одному временному промежутку (2020-2023 гг.) и написанных на одном языке (исключением могут являться, например, латинизированные названия компаний и пр.). Этот материал представляет собой несколько пользовательских подкорпусов, которые собирались по единым принципам.

3.2 Основные этапы обработки корпуса текстов скрытых сообществ

Многие исследования указывают на необходимость предобработки данных при построении тематических моделей независимо от используемого алгоритма [Чижик, 2021; Mamaev, Mamaeva, Axenova, 2022]. Первым этапом является *токенизация*, т.е. выделение единиц анализа. Для большинства языков мира такой единицей является слово. Второй этап заключается в приведении каждого выделенного токена в его каноническую форму, которая в дальнейшем будет проверяться на наличие в стоп-словаре.

Для организации единого лингвистического процесса предобработки использовалась библиотека *Stanza* для языка программирования *Python*. Она содержит инструменты, которые можно последовательно использовать для преобразования строковой формы текста в токены и леммы, а также получения необходимых морфосинтаксических характеристик [Qi et al., 2020]. Русскоязычная модель обучена на пяти наборах данных: SynTagRus, Taiga, GDS (Russian Universal Dependencies Treebank annotated by Google), PUD (Parallel Universal Dependencies) и Poetry. Так как большинство данных основаны на текстах онлайн-пространства, то платформа может использоваться для обработки текстов социальных сетей. Стоит отметить, что Stanza уже успешно зарекомендовала себя как библиотека для

предобработки текстов, функционирующих в вебе [Мамаев, 2022a; Mamaev, Mitrofanova, 2022a; Mamaev, Mitrofanova, 2022b; Sherstinova et al., 2023].

Для минимизации количества неверно распознанных токенов и приписанных им лемм процесс предобработки необходимо повторить несколько раз с последующим ручным анализом результатов. Ряд особенностей оформления постов социальных сетей, которые были описаны в главе 1, встретился в пользовательских текстах. Во-первых, во избежание блокировки некоторых постов со стороны администрации пользователи используют сочетание буквенных символов нескольких алфавитов со схожим графическим представлением. Например, при первичном визуальном анализе слова *«восстановление»* у одного из пользователей можно не заметить, что некоторые из символов кириллицы заменены на латиницу. При использовании специальных сервисов, один из которых – «Поиск кириллицы в латинице»[43], выявлено, что *«во**сс**тановлени**е**»* имеет три латинских символа, которые по написанию не отличимы от кириллических (см. выделение). Подобные замены были характерны для многих постов пользователя, поэтому слова со схожими комбинациями были добавлены в стоп-словарь. У других пользователей встречались и иные комбинации: кириллица и эмотиконы (*«сцене* ⭐*»*), кириллица и знаки пунктуации (*« : тарасова»*) и пр. Также из-за невозможности корректного отражения некоторых эмотиконов в стационарных текстовых редакторах и с целью минимизации шума при создании тематических моделей были удалены эмодзи, данный подход находит отражение в зарубежных работах. Так, в [Yang, Zhang, 2018, p. 528] для корпуса Twitter используются посты, из которых удалены фото- и видеоматериалы, а также эмотиконы.

После лемматизации нужно удалить некоторые токены с помощью стоп-словаря, созданного на основе «Нового частотного словаря русской лексики» О.Н. Ляшевской и С.А. Шарова[44] и Национального корпуса русского языка[45], в который

[43] http://invitemsg.com/cyrillic_search.php
[44] http://dict.ruslang.ru/freq.php
[45] https://ruscorpora.ru/

вошли лексические единицы, создающие информационный шум, поскольку они обладают широкой семантикой: предлоги, союзы, частицы, междометия и т.д. При повторных проверках в словарь вносились уже вручную выявленные нерелевантные единицы: обсценная лексика, ошибочно распознанные слова и пр. Итоговое количество стоп-слов для данного корпуса равняется 21400.

С точки зрения лексико-семантической организации некоторые языковые единицы (например, фразеологизмы) не могут быть разбиты на однословные [Нокель, Лукашевич, 2015], поэтому итоговым этапом стало автоматическое выделение биграмм в корпусе. Для каждого автора принято решение добавить в корпус двусложные сочетания, которые встречаются в посте более двух раз. С помощью библиотеки *gensim* и модуля *Phrases* извлекались двусловные сочетания. Символ *«_»* превратил униграммы в лексические конструкции: *«допустить_использование»*, *«психологический_помощь»*, *«главный_архитектор»*, *«территориальный_зона»* и т.д.

Таким образом, после обработки всех постов объем корпуса сократился почти в три раза.

3.3 Тематическое моделирование: обоснование выбора алгоритма и процедура построения

В области обработки текста стандартные процедуры тематического моделирования позволяют извлекать значимые тематические наборы слов как из структурированных, так и из неструктурированных данных [Bianchi et al., 2020]. Существующие вероятностные и алгебраические тематические модели не учитывают контекст, а сами тематические множества описывают темы широкой семантики, а не узконаправленные темы, характерные для данного автора [Mamaev, Mitrofanova, 2022]. В настоящее время предобученные языковые модели, такие как *BERT* или *Embeddings from Language Model* (*ELMo*), восполняют этот пробел, поэтому они используются в области семантической компрессии текстов. Один из вариантов контекстуализированного тематического моделирования представлен алгоритмом BERTopic, который представляет собой подход, использующий модели семейства Transformers и $c-TF-IDF$ для создания плотных кластеров,

которые соответствуют интерпретируемым темам, что позволяет сохранять важные слова в описаниях тем [Grootendorst, 2020]. По сравнению со стандартными вероятностными алгоритмами тематического моделирования, использование контекстуализированных моделей имеет ряд особенностей. В частности, BERT помогает сохранить векторные представления слов, учитывая контекст, что делает его более чувствительным к полисемантам [Митрофанова, Атугодаге, 2023]. Также, по заверениям авторов исследования [Egger, Yu, 2022], BERTopic показывает высокую результативность на разножанровых корпусах.

Алгоритм BERTopic состоит из трех этапов: создание эмбеддингов документов, предсказание семантических кластеров и вывод темы из кластеров. Алгоритм $c-TF-IDF$ сравнивает важность лексических единиц с конкретным кластером и выявляет наиболее значимые лексические единицы в теме. Значение $c-TF-IDF$ рассчитывается по следующей формуле (3).

$$c-TF-IDF = \frac{f_i}{wd_i} \times log\frac{m}{\sum_j^n f_j} \quad (3)$$

Частота в формуле (3) для каждого слова f извлекается из каждого конкретного кластера i, а затем делится на общее количество лексических единиц wd кластера i. Это способ нормализации частоты слов в каждом кластере. Затем количество кластеров m делится на общую частоту слова f во всех кластерах. После создания представлений $c-TF-IDF$ пользователь получает набор лемм или словоформ (в зависимости от того, проведена ли нормализация текста), который характеризует набор текстов. Конечно, это не означает, что набор слов описывает связную тему. Чтобы повысить связность слов в темах, алгоритм использует максимальную предельную релевантность (*Maximal Marginal Relevance, MMR*), чтобы найти наиболее связные слова без слишком большого совпадения между самими словами. Это действие приводит к удалению слов, которые не относятся к определенной теме. Графическое представление архитектуры BERTopic представлено на рисунке 5.

Рисунок 5 — Архитектура BERTopic

Мультимодальность BERTopic, т.е. возможность создания тематической модели «не только по словам, но и по любым терминам другой модальности» [Смелик, Фильченков, 2016, с. 424], обеспечивается наличием многоязычной языковой модели (русский язык также учитывается), благодаря которой имеется возможность обрабатывать входные тексты, которые содержат слова из различных языков, интеграцией выделенных ранее биграмм, а также разделением исходного корпуса на авторские подкорпусы.

Для решения проблемы унификации вывода определенного количества слов-тематизаторов вручную был добавлен дополнительный фильтр. Поскольку библиотека BERTopic автоматически подбирает количество слов-тематизаторов, то нужно зафиксировать единый количественный показатель для всех результатов. Количество тем в настоящем эксперименте равно 10, поскольку стандартные библиотеки, работающие с LDA-моделями и их модификациями, выводят именно это количество лемм.

С учетом описанных особенностей алгоритма и выставленных фильтров для 704 пользователей только 376 были представлены в тематических моделях, что составляет 53% от общего корпуса. Подобное количество полученных моделей объясняется несколькими причинами. Во-первых, BERTopic применительно к корпусам направлен на снижение размерности и структурное обобщение, что проявляется в ранжировании авторов по значимости их влияния на корпус. Например, в стандартной процедуре автор-тематической моделирования,

реализованной в *gensim*, каждый пользователь получит хотя бы одну тему, но итоговая модель окажется избыточной и нерепрезентативной из-за пересечения множеств слов [Mamaev, Mitrofanova, 2022a; Mamaev, Mitrofanova, 2022b]. Во-вторых, стандартные процедуры фильтрации текстов стоп-словарем уменьшают объем корпуса за счет удаления «шумных» единиц, в результате чего обработанные пользовательские подкорпусы с малым количеством слов не обрабатываются алгоритмом. Если не включать в стандартные процедуры предобработки текстов фильтрацию стоп-словарем, то итоговые наборы тем будет сложно интерпретировать лингвистам-экспертам. Общее количество тематических моделей – 2217. В таблице 1 приведем ряд дискретных автор-тематических моделей по пользователям.

Таблица 1 — Примеры дискретных автор-тематических моделей BERTopic

ID пользователя	**Тема**	**Слова-тематизаторы**
1	**2**	**3**
135242	0	жизнь, любить, стол, работать, английский, ребята, день, умный, далеко, хороший
	1	место, турнир, танец, категория, стандарт, класс, латина, юниор, легион, клуб
	2	учитель, преподаватель, учебный, помощь, нынешний, ученик_бывший, жизнь, учиться, друг, километр
77955	0	предложение, выделять, союз, ставить, оборот, слово, нужный, писать, правило, язык
	1	билет, концерт, музыкальный, новый, день, альбом, место, клуб, песня, выпустить
33970	0	флот, фрегат, морской, ракета, ракетный, адмирал, горшков, цель, балтийский, море
	1	завод, оборудование, позволить, кремний, российский, тонна, несколько, насос, скважина, создать
	2	сердце, взгляд, судьба, встреча, суббота, слышать, безнадежный, начаться, подобный, погнать

Продолжение таблицы 1

1	2	3
25256	0	спичка, секрет, ян, больно, тимур, чиркнуть, мкр, первый, загадка, пробовать
	1	дождь, солнце, трава, ветер, ущелье, река, метр, туман, холодный, делать
	2	душа, жизнь, мир, бог, жить, чужой, сердце, переставать, страшный, надежда
	3	осень, время, вечер, приключение, минута, октябрь, вечность, следующий, лето, саша
	4	любить, понимать, человек, совет, главное, друг, отбить, понять, помощь, любопытство
	5	картошка, горячий, бабушка, шоколад, шоколадка, банк, вкусный, пробовать, разный, продать

Проведем сравнение контекстуализированных тем с традиционными темами, которые были сгенерированы алгоритмом LDA. В качестве лингвистического инструмента выступила онлайн-среда jsLDA: In-browser topic modeling[46]. Это онлайн-платформа тематического моделирования, основную цель которой можно сформулировать в трех постулатах: а) упростить создание тематических моделей непосредственно через веб-браузер для специалистов гуманитарной сферы деятельности, б) продемонстрировать потенциал статистических вычислений в Javascript, в) обеспечить более тесную интеграцию между веб-визуализациями и итоговыми моделями. В систему были загружены заранее предобработанные и отфильтрованные в *Stanza* посты пользователей, в диапазоне от 50 до 100 итераций был запущен алгоритм. Результаты представлены в таблице 2.

[46] https://mimno.infosci.cornell.edu/jsLDA/jslda.html

Таблица 2 — Примеры LDA-тематических моделей

ID пользователя	**Тема**	**Слова-тематизаторы**
1	**2**	**3**
135242	0	жизнь, мама, английский, время, день, разный, учитель, сбор, преподаватель, стол
	1	всероссийский, мероприятие, работа, история, забота, киселева, конкурс, строка, умный, парень
	2	категория, новый, турнир, юниора, санкт-петербург, открытый, герасимов, поздравлять, дарья, участник
	3	место, стандарт, класс, латина, клуб, легион, тск, екатерина, чемпионат, пара
	4	танец, ребята, танцор, турнир, человек, результат, хотеться, ситуация, танцевальный, прекрасный
77955	0	выделять, ставить, знак, оборот, язык, обстоятельство, частица, тир, местоимение, тип
	1	писать, вопрос, начало, запятая, однородный, концерт, билет, фразеологизм, область, весна
	2	предложение, союз, слово, нужный, русский, правило, часть, определение, сравнение, вопросительный
33970	0	россия, научный, рота, новый, кремний, завод, вмф, призывник, подразделение, несколько
	1	флот, фрегат, ракета, морской, адмирал, горшков, успешно, пресс-служба, балтийский, море
	2	ракетный, цель, зенитный, тысяча, специалист, радиотехнический, служба, два, полимент, экипаж
	3	использовать, российский, первый, позволить, сердце, оборудование, насос, тонна, встреча, взгляд
25256	0	жить, дождь, ветер, понравиться, земля, глаз, настоящий, господь, плечо, глубина
	1	человек, душа, новый, жизнь, живой, любить, солнце трава, помнить, нужный
	2	друг, бояться, история, начало, ребята, смотреть, работа, спрашивать, свобода, море
	3	разный, горячий, большой, жизнь, человек, взорваться, рассыпать, тоска, чужой, верить

1	2	3
25256	4	первый, делать, вертолет, дорога, бог, неожиданный, пробовать, главное, хороший, путь
	5	время, мир, день, рука, прийти, вечер, фрэнк, идея, километр, гора

Несмотря на выдачу большего количества тематических моделей, необходимо перечислить ряд недостатков полученных LDA-моделей. Во-первых, несмотря на насыщение корпуса биграммами, они не были включены в итоговые темы в отличие от BERT-моделей (см. *«ученик_бывший»* у пользователя 135242). Во-вторых, LDA оказывается чувствителен к размеру пользовательского подкорпуса, поскольку итоговые выдачи могут сочетать в себе несколько тем, в то самое время как тематические модели BERT однородны по составу. Например, тема 1 у пользователя 77955 (см. таблицу 2) включает в себя слова-тематизаторы, посвященные тематике русского языка и концертной тематике, в выделенных BERT-темах для этого же пользователя (см. таблицу 1) подобного пересечения тем в рамках одного множества слов не наблюдается. Наконец, в темах LDA преобладают лексические единицы с широким значением, под которыми понимается «более широкое качество связи между десигнатором (формой слова) и десигнатом (содержанием, значением слова), обеспечивающее нестесненную широту семантического варьирования» [Никитин, 2005]. Большое количество таких единиц в тематической модели вызовет проблемы при выборе верной метки темы. Так, в темах 4 и 5 у пользователя 25256 (см. таблицу 2) встречаются такие многозначные слова, как *«путь»* (12 значений в лексикографическом ресурсе Викисловарь[47] по состоянию на 01.02.2024), *«хороший»* (7 значений в лексикографическом ресурсе Викисловарь по состоянию на 01.02.2024), *«время»* (9 значений в лексикографическом ресурсе Викисловарь по состоянию на 01.02.2024) и пр. Таким образом, в рамках настоящего эксперимента используются контекстуализированные наборы тем.

Для автоматически сформированных слов-тематизаторов нужно провести процедуру унификации для удобства создания модели скрытых сообществ. В этом

[47] https://ru.wiktionary.org/wiki

случае размечаются темы, т.е. подбираются общие слова и фразы, которые охватывают содержание текста, подвергнутого семантической компрессии.

3.4 Процедура ручного аннотирования тематических моделей

Существенным недостатком тематических моделей является невозможность автоматически обобщить содержание тем текста. В оригинальной форме алгоритм просто упорядочивает слова внутри темы по вероятностям и присваивает каждой теме номер, не предоставляя средств для автоматического назначения меток тем. На сегодняшний день существуют два подхода: ручное аннотирование с привлечением экспертов и автоматическое аннотирование с привлечением внешних/внутренних источников знаний.

Одна из первых работ, посвященных автоматическому назначению меток тем, появилась в 2007 году [Mei, Shen, Zhai, 2007]. Задача заключается в присвоении семантических меток из корпуса множеству слов, которые полностью бы отражали тему. Исследователи предложили два метода для автоматического назначения меток тем. Первый метод заключался в извлечении фраз, а второй – в получении n-грамм. В первом методе под фразой понимаются единицы, обладающие грамматической и смысловой цельнооформленностью и не требующие корректировки с точки зрения языка. Второй метод характеризуется независимостью от обучаемой области, однако, с лингвистической точки зрения, он не всегда может обеспечить семантическую целостность, и грамматические связи могут быть нарушены (потеря согласовательных связей в результате лемматизации).

В дальнейшем получили развитие и другие методы, в том числе с использованием внешних источников. В [Bhatia, Lau, Baldwin, 2016] кандидаты в метки генерировались на основе англоязычной Википедии, после чего осуществлялось ранжирование при помощи косинусного сходства. В исследовании [Allahyari et al., 2017] была предложена модель LDA, дополненная понятиями из формальной онтологии, а для поиска наиболее значимых меток использовался семантический граф. В статье [Mitrofanova et al., 2021b] одним из способов интерпретации тематических моделей является вычленение хэштегов из постов

корпуса Pikabu. В работе [Mitrofanova et al., 2021a] предложены два подхода к назначению меток тем специализированных корпусах русского языка: в рамках первого метода извлечение меток-кандидатов осуществлялось из поисковой системы Яндекс, а второй – из Википедии с применением процедуры эксплицитного семантического анализа (*explicit semantic analysis*). Результаты показали, что первый алгоритм предсказывает метки, но не всегда корректно связывает их со словами-тематизаторами. Второй алгоритм выделяет метки высокого уровня абстракции. Оптимальным решением, позволяющим автоматически обобщить содержание тем в текстах с учетом контекста из внешних источников, является комбинация этих методов. Наконец, могут использоваться и диалоговые системы, обученные на бóльших данных, например, ChatGPT [Тен, 2023], однако итоговые метки тем оказываются чувствительны к временным параметрам: если тема описывает событие, информации о котором нет в данных, то метка будет стремиться к максимальному обобщению всех лемм в теме.

Несмотря на расширение круга автоматических методов, наиболее надежным является ручной способ аннотирования тем, в пользу которого приводятся следующие аргументы. Во-первых, для разметки моделей возможно выбрать одну из существующих классификаций текстов, которые представлены в общедоступных корпусах. Во-вторых, ручная разметка позволяет исследователю иметь больший контроль над процессом и полученными результатами. В-третьих, применение оценок согласованности аннотаторов (каппы Флейса или каппы Коэна) позволяет снизить уровень субъективной оценки, они активно используются в задачах, которые решаются с помощью тематического моделирования (см. [Hagen, 2018; Palese, Piccoli, 2020]). При обращении к каппам необходимо учитывать предметные области. В гуманитарных областях или при проведении оценки какого-то повседневного опроса результат согласованности в 40% уже можно считать достоверным, согласно [McHugh, 2012] и информации с сайта Statistics How To[48], что нехарактерно для технических или медицинских областей. В качестве ведущей

[48] https://www.statisticshowto.com/cohens-kappa-statistic/

классификации для ручного аннотирования тем была выбрана классификация НКРЯ[49], так как она представлена в виде обобщенных предметных областей (тем-гиперонимов), что минимизирует их возможное пересечение [Савчук, 2005]. В данной классификации содержится 55 тем (по состоянию на март 2023 г.), при этом ведущими темами являются «политика и общественная жизнь» и «частная жизнь».

Разметка производилась автором работы в период с марта по апрель 2023 г. согласно данной классификации, которая была незначительно расширена в зависимости от анализируемых слов-тематизаторов. Было введено две дополнительные темы: «журналистика», описывающая труд и жизнь работников прессы, а также «рабочий процесс», в рамках которой представлен непосредственный процесс создания чего-либо. Такие темы были присвоены 13 моделям из 2217, что составляет 0.0059%.

Оценка тем производилась силами двух лингвистов-экспертов в мае 2023 г., перед которыми было поставлено следующее задание: ознакомиться со словами-тематизаторами и предлагаемой для них темой; при согласии с предложенной темой поставить 1, а при несогласии – 0. По возможности лингвистам предоставлялась возможность предложить свою тему на основе ведущей классификации. Результаты опроса были представлены в виде матрицы согласованности (таблица 3).

Таблица 3 — Матрица согласованности для расчета каппы Коэна

	Аннотатор 2		**Итог**
Аннотатор 1	***1***	***0***	
1	1771	146	1917
0	157	143	300
Итог	1928	289	**2217**

Значение каппы Коэна рассчитывалось по следующей формуле (4).

$$k = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

[49] https://ruscorpora.ru/stats

Значение p_o является относительным согласием между лингвистами, это доля от общего числа оценок, на которые оба лингвиста дали либо положительный, либо отрицательный ответ. Значение p_e – вероятность того, что согласие лингвистов носит случайный характер. В результате вычислений в среде Excel итоговое значение каппы Коэна равняется 40.86%, что удовлетворяет минимальным требованиям при проведении оценки [McHugh, 2012]. Для ручной разметки преодоления порога в 40% также достаточно для того, чтобы в исследовательском наборе данных появились альтернативные метки тем.

В отличие от методов автоматического назначения тем [Mamaev, Mitrofanova, 2020a], при ручном аннотировании с опорой на классификацию метки не нужно унифицировать. Следующим этапом работы стало создание модели скрытых сообществ и ее характеристика.

3.5 Итоговая модель скрытых сообществ и ее формальные характеристики

Разрабатываемую модель сообществ необходимо представить в виде графа, узлами которого выступят ID пользователей, а ребрами – скрытые связи между узлами. В качестве программного обеспечения выступил комплекс из двух инструментов. Первый из них – *Easy Linavis*[50], его разработали для создания графа персонажей в художественном произведении. Такой подход можно спроецировать и на реальные сообщества. Оформление кода сводится к следующим постулатам: с помощью символа # идет обозначение тематической группы, далее на каждой отдельной строчке идет условное обозначение пользователя. Пример реализации программы представлен на рисунке 6.

[50] https://ezlinavis.dracor.org/

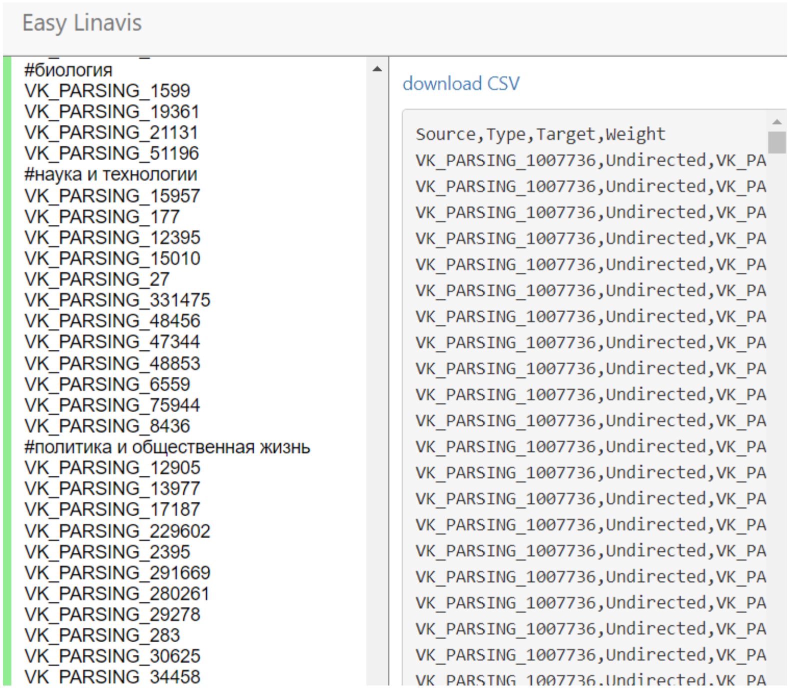

Рисунок 6 — Организация меток тем и пользователей для создания графа

В дальнейшем необходимо скачать CSV-файл, требующийся для отрисовки итогового графа в приложении *Gephi*[51], программном обеспечении с открытым исходным кодом для работы с сетями различной сложности. *Gephi* предоставляет пользователю возможность исследовать и анализировать сложные структуры, обнаруживать закономерности и взаимосвязи в сетях. Этот инструмент важен с точки зрения визуализации данных. Итоговый граф сообществ представлен на рисунке 7.

[51] https://gephi.org/

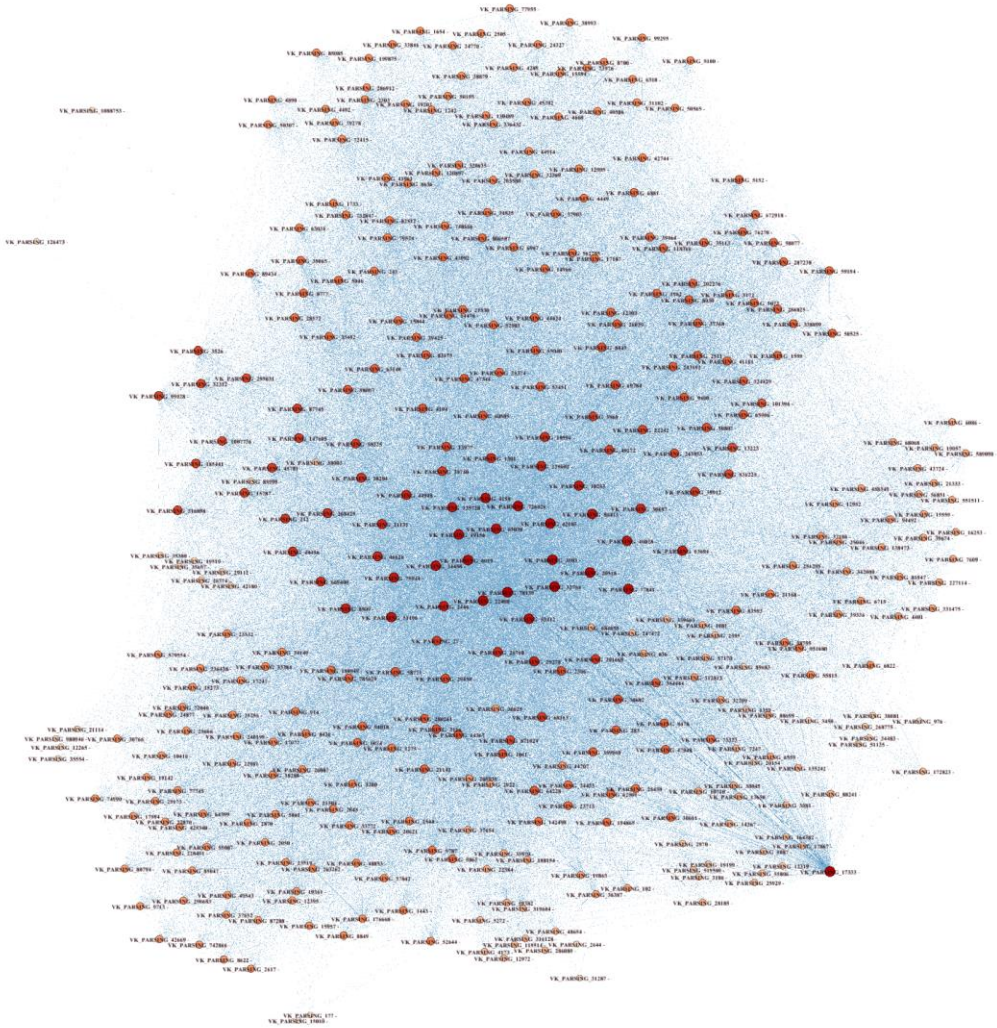

Рисунок 7 — Модель скрытых сообществ на основании общности тематических компонентов пользовательских постов

Для полученного графа можно дать оценку с двух точек зрения: формальной и социально-демографической [Мамаев, Митрофанова, 2024]. Формальные характеристики были получены автоматически при построении графа в *Gephi*.

Таблица 4 — Формальные характеристики графа

Параметр	**Значение**
1	**2**
Количество узлов	376
Количество ребер	34507
Плотность графа	0.489
Диаметр графа	3

Продолжение таблицы 4

1	2
Тип графа	неориентированный
Средний коэффициент кластеризации графа	0.823
Модулярность	0.167
Предполагаемое количество сообществ на основании расчета модуляции	4

Для автоматически построенного графа существует два подхода для оценки качества потенциально выделенных сообществ. В первом случае неизвестно истинное разбиение на сообщества. Такая ситуация встречается при работе с большими данными. В этом случае для оценки качества часто используется значение модулярности [Fortunato, 2010]. Во втором случае истинное разбиение известно. Такой подход применим для графа знакомств друзей пользователя в социальной сети (так называемый эго-граф), в котором пользователь самостоятельно может разделить всех друзей на группы. В следующих разделах будет описан второй сценарий, когда вручную будет вычислено реальное количество сообществ. На данном этапе вводится предположение, что реальное количество сообществ в модели нельзя вычислить.

Так, полученный средний коэффициент кластеризации показывает большое количество потенциальных групп внутри сети, т.е. она является неоднородной. Плотность графа 0.489 указывает на средний уровень связанности ее участников (более 300). Модулярность как скалярная величина в диапазоне $[-1;\ 1]$ указывает на то, насколько плотность связей внутри сообщества больше плотности связей между сообщества при полученном разбиении сети на сообщества. Показатель, приближающийся к нулю, позволяет утверждать, что различия в плотностях в группах и между ними явно не выражена. Согласно [Чеповский, 2023], полученные параметры действительно характеризуют структуру социальных сетей, которые имеют ряд особенностей: «маленький диаметр графа (эффект «малого мира»), высокие значения кластерного коэффициента (эффект «транзитивности»)».

С помощью библиотеки *vk_api* извлечены базовые социально-демографические параметры пользователей скрытых сообществ. Так, основная

часть пользователей, входящих в итоговую модель скрытых сообществ, проживает в городах-мегаполисах, а оставшийся хвост распределения характеризует точечные активности из других городов РФ и мира в целом (рис. 8).

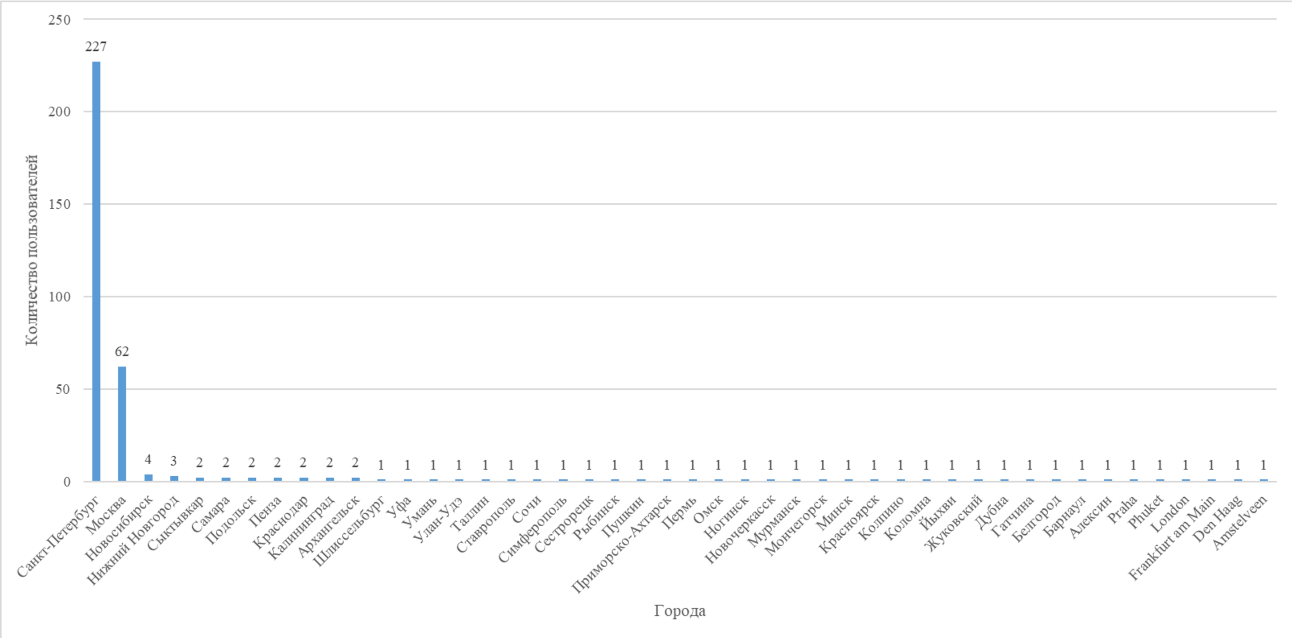

Рисунок 8 — Распределение пользователей по городам

Необходимо отметить, что место проживания, как и некоторые другие параметры, заполняются не всеми пользователями, что является потенциальной проблемой для социологов, которые работают с большими данными. Ее решением может стать постоянный мониторинг обновлений, совершенствование программ обработки данных, а также анализ только заполненных полей. В настоящем эксперименте нет задачи отслеживания динамики развития модели сообществ, в том числе с лингвистической точки зрения, был выбран статический «информационный портрет» корпуса. Всего 345 пользователей указали место проживания.

Данные, представленные на рисунке 9, частично подтверждают результаты исследований, проведенные в начале 2010-х гг. [Lenhart et al., 2010; Rainie, Lenhart, Smith, 2012]: взрослые от 30 до 49 лет являются активными пользователями социальных сетей. 157 пользователей указали дату рождения с годом, 163 пользователя указали день и месяц рождения, 53 пользователя не указали никаких данных. Еще три пользователя указали полную дату рождения, однако в подсчет

она не включалась, поскольку дата рождения не соотносилась с визуальной информацией, представленной на странице (например, фотографии и картинки, которые не характерны для данного возрастного периода): это пользователи с датами 11.11.1918, 26.11.1925 и 11.11.1928. В частности, пользователь, который по данным ВКонтакте родился в 1925 году, выкладывает актуальные фотографии, оставляет посты, один из них является благодарностью за поздравление с днем рождения. К посту прикреплена песня Сектора Газа «Мне снова 30 лет».

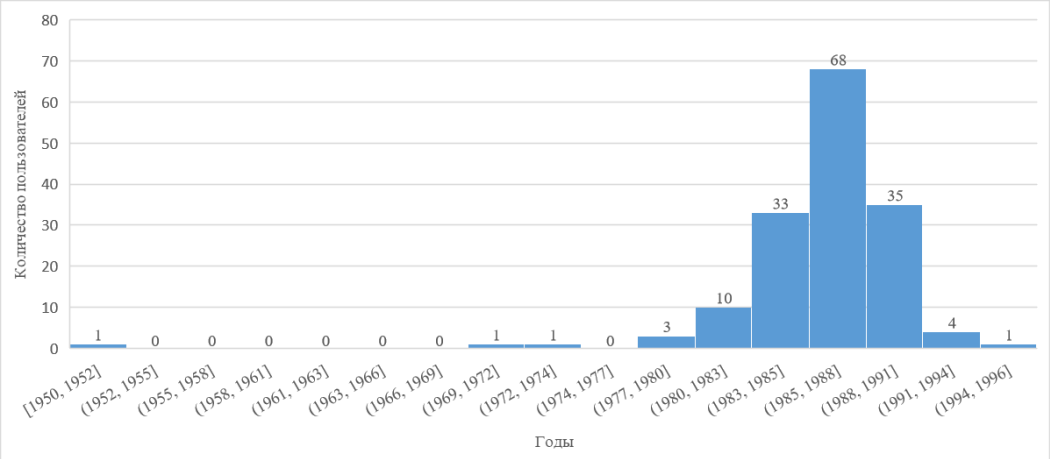

Рисунок 9 — Распределение пользователей по годам рождения

Несмотря на то, что часть пользователей не вошла в итоговую модель, распределение по полу осталось приблизительно равномерным (рис. 10). Этот параметр оказался единственным, который заполнен у всех пользователей.

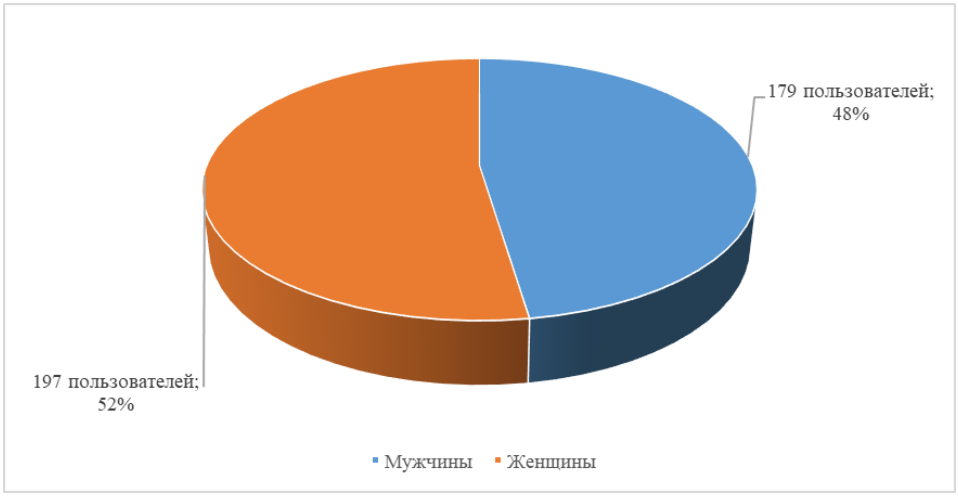

Рисунок 10 — Распределение пользователей по полу

Наибольшую проблему унификации данных представляли интересы и образование пользователей, так как одни и те же именованные сущности (названия

вузов) и лексические единицы, обозначающие хобби, могут быть представлены по-разному: ср. *БГТУ «Военмех» им. Д.Ф. Устинова/Балтийский государственный технический университет «Военмех» им. Д.Ф. Устинова* и пр., *snowboard/сноуборд* и пр. Некоторые пользователи, указывая интересы, пользовались конструкциями широкой семантики. Например, пользователь с ID 6559 указал: «Все самое интересное и увлекательное». Поэтому для данных параметров указано лишь наличие и отсутствие значений (рис. 11 и 12).

Рисунок 11 — Распределение пользователей по указанному высшему образованию

Рисунок 12 — Распределение пользователей по указанным интересам

Наконец, стоит охарактеризовать основные положения по реальному количеству скрытых сообществ в настоящей модели. Среди 376 пользователей в итоге было выявлено 34 скрытых сообществ (рис. 13), описывающие как

широкопрофильные, так и узконаправленные тематики. Формальными методами ранее было выделено 4 сообщества. Количественная разница объясняется тем, что структура скрытых сообществ обладает менее устойчивыми связями, которые лучше выявляются комбинациями автоматических методов и ручных подсчетов.

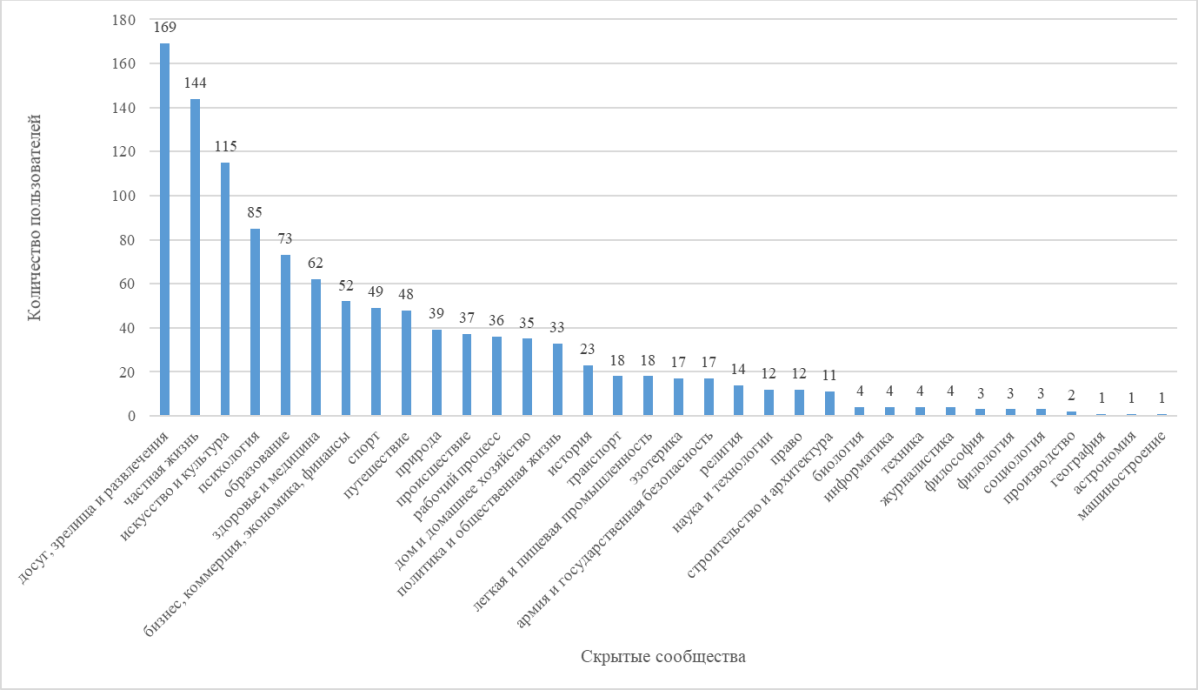

Рисунок 13 — Количество пользователей в скрытых сообществах

В предыдущих исследованиях [Mamaev, Mitrofanova, 2020a; Mamaev, Mitrofanova, 2020b] тема здоровья была ведущей в связи с распространением COVID-19. Для текущей модели характерно большое преобладание пользователей в сообществах повседневного характера, несмотря на сложную геополитическую обстановку в мире.

На рисунке 14 представлено соотношение пользователей по количеству сообществ, в которых они состоят. Около 57% пользователей состоят в двух-четырех скрытых сообществах, пик распределения характеризует многостороннюю заинтересованность пользователей различными областями, а хвост указывает на приверженность единому интересу.

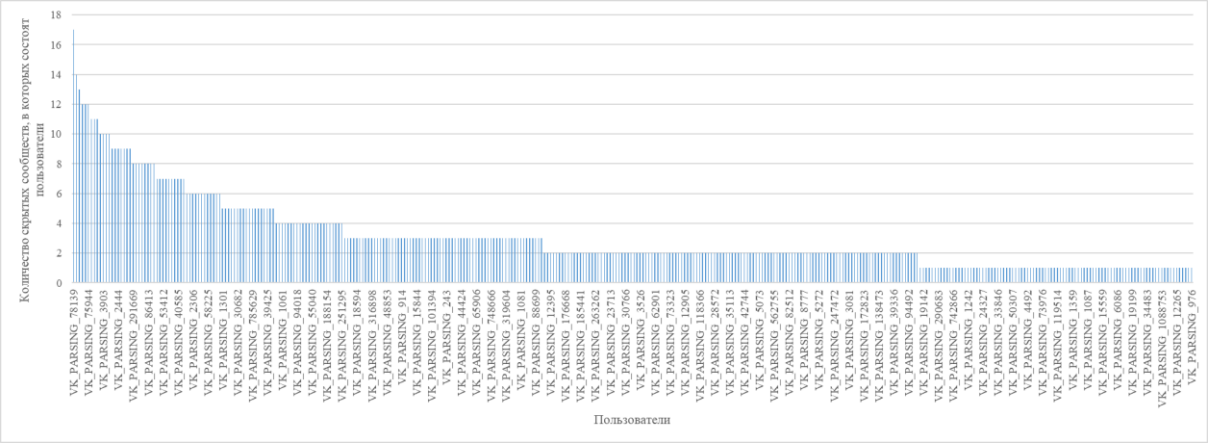

Рисунок 14 — Соотношение количества пользователей по количеству скрытых сообществ

Все полученные данные следует воспринимать не как отражение реальной картины мира, а лишь некоторый цифровой след, которые оставляют пользователи – *‘цифровые личности’* (*ЦЛ*) в социальных сетях. По мнению Е.В. Чернявской, сетевое пространство открывает возможности для широкого распространения информации о различных социальных типизациях, формах идентичности и социальных ролях. Возникли новые способы формирования цифровой личности, она становится стилизуемой в соответствии с уже существующими цифровыми образцами [Чернявская, 2020, с. 10]. Цифровая личность проявляется как личность, приобретающая ряд нехарактерных для традиционного дискурса привилегий, таких как игнорирование адресата, множественная адресация, отсутствие физической дистанции между собеседниками, использование языковых ресурсов для воздействия на сознание собеседника [Попова, 2019]. Таким образом, цифровую личность можно определить как «всю совокупность информации, оставленной о себе человеком в электронном пространстве – в своих аккаунтах и постах, поисковых и иных запросах, подписках, историях просмотров, оценках информационных ресурсов («лайках»), публикации персональных данных, научном и художественном творчестве, следах от оплаты товаров и услуг банковскими картами и т. д. Согласно закону, для научного исследования доступна только та часть ЦЛ, которая находится в открытых источниках…» [Лихачева, 2020, с. 246]. При этом, по мнению А.В. Бухановского[52], цифровые личности делятся на

[52] https://news.itmo.ru/ru/news/8695/

две категории. К первой категории относятся абстрактные синтетические личности, для которых не важен факт существования реального человека. Под эту категорию относятся, например, фан-страницы живых или мертвых знаменитостей. Вторая категория, ассоциированная, развивает особенности реального человека-прототипа. Таким образом, в рамках исследования основой фокус направлен на цифровом образе скрытых сообществ, опускается вопрос соотношения пользователей сообществ с реальными людьми.

3.6 Отбор признаков для проведения процедуры лингвистического профилирования

Профилирование автора – это процесс выявления взаимосвязи личных характеристик автора (пол, возраст, профессия…) и его особенностей создания текстов на основании созданного им корпуса [Литвинова, 2013; Halteren, 2004; Daelemans et al., 2019], данный подход является основным аппаратом для проведения судебных экспертиз, психологических исследований и пр. С развитием вычислительных мощностей появились разнообразные способы профилирования текстов, что привело к появлению отдельных задач в этой области: определение родного языка [Cimino et al., 2017], определение жанра [Paltridge, 1994], атрибуция автора [Gamon, 2004] и пр. Несмотря на большие достижения в области профилирования, многие исследователи до сих пор обсуждают ряд вопросов, один из которых – выбор необходимых для профилирования параметров [Nini, 2014]. В работах [Dell’Orletta, Montemagni, Venturi, 2013; Brunato et al., 2020] отмечается, что основные характеристики, рассматриваемые при профилировании, сводятся к длине языковых конструкций в необработанном тексте, распределениям по употреблению основных частей речи и др. Основные характеристики представлены в таблице 5.

Таблица 5 — Лингвистические параметры профилирования текста

<table>
<tr><th>Общие параметры исходного текста</th><th>Лексические параметры</th><th>Морфологические параметры</th><th>Синтаксические параметры</th></tr>
<tr><td>Длина документа – среднее количество предложений в документе.</td><td rowspan="4">Соотношение числа разных лексических единиц к общему числу лексических единиц в исследуемом отрывке (Type-token ration). Длину исследуемого отрывка устанавливает исследователь.</td><td>Лексическая плотность – отношение количества слов определенной части речи ко всем словам исследуемого текста.</td><td>Особенности глагольного предиката: изучение левых и правых зависимостей, проверка на наличие или отсутствие подлежащего и пр.</td></tr>
<tr><td>Длина предложения – среднее количество слов в предложении.</td><td rowspan="3">Распределение грамматических категорий – выявление вероятности появления слова определенной части речи в тексте.</td><td>Взаимосвязь главных и придаточных частей предложения.</td></tr>
<tr><td rowspan="3">Длина слова – среднее количество символов в слове.</td><td>Глубина синтаксических деревьев.</td></tr>
<tr><td>Порядок слов.</td></tr>
<tr><td>Средняя длина как главных, так и зависимых частей предложения.</td></tr>
</table>

Согласно [Litvinova, 2014], первые методы языкового анализа авторских данных (так называемые *content-based*) были сосредоточены на изучении отдельных лексико-семантических классов (например, эмоционально окрашенные единицы) и установлении их взаимосвязи с типами личности человека. Однако такой подход имеет ряд недостатков: в частности, группы слов отбираются на основе субъективных представлений ученого, которые некоторым образом зависят от его культуры.

Второй подход основан на извлечении сведений о других языковых уровнях, не только о лексическом. Значительный прогресс в этой области стал возможен благодаря созданию достаточно точных инструментов анализа текста для различных языков [Daelemans, 2013]. Эти программы используют различные функции для оценки общей лингвистической сложности текстовых массивов. Например, библиотека *Stylo* для языка программирования R [Eder, Rybicki, Kestemont, 2016] позволяет проводить всесторонние стилометрические исследования. Этот пакет извлекает статистически значимые *n*-граммы на уровне токенов и символов, которые могут быть автоматически получены без использования дополнительных средств, зависящих от одного конкретного языка.

Похожий подход наблюдается в приложении *Orange* [Demšar et al., 2013] при комбинации инструментов *Textable* и *Text*. Они позволяют извлечь базовые лингвостатистические данные: средняя длина сегментов, количество слов и символов, соотношение гласных и согласных, читабельность текста по индексу LIX и т.д.

Синтаксические анализаторы *TAASSC* [Kyle, 2016] и *L2SCA* [Lu, 2010] оценивают грамматическую сложность словосочетаний и предложений. Эти инструменты выделяются среди прочих тем, что они были разработаны для оценки текстов, которые изначально не были созданы носителями языка, то есть их практическая значимость заключается в улучшении методики преподавания иностранных языков при выявлении синтаксически сложных или аграмматических конструкций.

Profiling-UD [Brunato et al., 2020] позволяет извлекать более 130 признаков (лексические, морфосинтаксические и пр.) из исследуемого текста. В основе инструмента лежит подход универсальных зависимостей (Universal Dependencies), он поддерживает 59 языков.

Переходя к отдельным исследованиям профилирования авторов, следует отметить, что пол – наиболее частый параметр для изучения [Cheng et al. 2009; Cheng, Chandramouli, Subbalakshmi, 2011; Marquardt et al., 2014], однако математический аппарат, предлагаемый авторами, применим к англоязычным датасетам. Среди отечественных научных групп проблемой профилирования занимается Научно-исследовательская лаборатория компьютерной семасиологии, которая на протяжении многих лет выявляет корреляции между языковыми параметрами и отдельными параметрами личности человека. Например, в работе [Litvinova et al., 2018] на материале русских письменных текстах описан эксперимент по выявлению различий в текстах, составленных мужчинами, и текстах, составленных женщинами. Авторы также разрабатывают математическую модель для определения пола авторов текстов с использованием только высокочастотных параметров текста, независимых от тематики. Анализ показал, что в текстах, написанных на русском языке мужчинами, по сравнению с текстами, написанными женщинами, выше индекс лексического разнообразия и доля

предлогов и местоименных прилагательных. Тексты мужчин оказались стилистически более формальными, чем тексты женщин.

В другом исследовании научной группы [Litvinova, Sboev, Panicheva, 2018] сделан акцент на изучении возрастных характеристик блогеров. Ученые собрали корпус постов, опираясь на русскоязычный сегмент социальной сети LiveJournal. Данная текстовая коллекция представлена более чем 1200 авторами, она сбалансирована как с точки зрения пола, так и с точки зрения возрастных характеристик. Возрастная классификация проводилась в трех группах (от 20 до 30 лет, от 30 до 40 лет, от 40 до 50 лет) и с использованием лексических единиц различного уровня (униграмм, биграмм и т.д.). Было установлено, что комбинация лексических единиц с частеречной разметкой и уровнем лексического разнообразия почти не улучшает классификационные модели. Полученные модели классификации и регрессии лучше работают для текстов, авторство которых приписывается лицам женского пола, чем для текстов мужчин.

В рамках исследования не представляется возможным провести полномасштабную оценку взаимосвязи социопсихологических характеристик пользователей с лингвистическими параметрами порождаемых ими текстов. Во-первых, для выявления психологических параметров необходимо провести ряд специализированных тестов с реальными личностями. Учитывая, что за цифровым образом пользователя может стоять совершенно другой человек, а также тот факт, что часть пользователей может отказаться от эксперимента, итоговые данные могут оказаться несостоятельными. Во-вторых, как ранее отмечалось, признаки возраста, образования и др. (за исключением параметра пола) указываются не всеми пользователями. По этим причинам в рамках настоящей работы будут выявлены внутритекстовые статистические связи. В настоящем эксперименте вводится определение ‘*лингвистический профиль пользователей скрытого сообщества’* – набора языковых признаков, которые выявляются на основе текстовых массивов, составленных участниками той или иной группы [Мамаев, 2024b]. Такие профили потенциально можно использовать при автоматической рубрикации текстов постов

социальных сетей. В работе рассмотрено три вида корреляций на морфологическом, синтаксическом и лексическом уровнях.

Первое предположение – предположение о корреляции имен существительных и глаголов и их модификаторов (имен прилагательных и наречий соответственно), что описано в [Бодрова, Тукмакова, 2012; Тукмакова, 2020; Hengeveld et al., 2007]. Данный вопрос затрагивался А.А. Потебней в труде «Из записок по русской грамматике»: в работе отмечена близость имен существительных и имен прилагательных, которая берет начало из единства категории имени: «Существительные и прилагательные в тесном смысле …, будучи близких к глаголу, еще более близки между собою» [Потебня, 1958]. Эту близость можно оценить и количественно. Например, сильная положительная связь между существительным и прилагательным позволяет передать не только понятийное содержание объекта, но и дополнительную информацию о его характеристиках. При сильной положительной корреляции глагола и наречия можно наблюдать авторскую тенденцию к расширению описания самого действия. Отрицательные корреляции будут свидетельствовать об отклоняющихся от стандартных тенденций идиостиля самой тематической группы.

Второе предположение – предположение о синтаксической сложности текстов тематической группы. Было решено выбрать два вида корреляций – корреляция длины предложения и количества предложных конструкций, а также корреляция длины предложения и длины связей зависимостей. В работе [Конюшкевич, 2013] на материале белорусского языка показано, что корреляция предикативной части сложного предложения и соответствующей именной синтаксемы с предлогом обладает сильной связью. Наряду с этим можно рассчитать корреляцию длины предложений и длины связей зависимостей, которая описана на материале индоевропейских и азиатских языков [Liu, 2008; Jiang, Liu, 2015]. Более длинные предложения и сложные структуры зависимостей затрудняют понимание тематического поста, что влияет на его обсуждение и, как следствие, становится причиной коммуникативной неудачи. Авторы постов смогут принимать обоснованные решения о стиле своих дальнейших сообщений, чтобы

оптимизировать читабельность текста и повысить вовлеченность целевой аудитории. Также данная корреляция может стать одной из качественных лингвистических процедур выявления ключевых влиятельных лиц и тенденций в социальных сетях наравне с формальными показателями (реакции, просмотры и пр.). Сообщения в социальных сетях, которые содержат более длинные предложения и более длинные связи зависимости, с большей вероятностью будут использоваться лидерами мнений и влиятельными лицами в определенной нише.

Наконец, третья предположение – предположение о связи лексического разнообразия и лексической плотности текстов тематических групп скрытых сообществ. В [Стрельников, Воробьева, 2022] выдвигается предположение о зависимости критериев информационной насыщенности (лексической плотности) от метрик лексической насыщенности (стандартный ttr). В качестве исследовательского корпуса взяты тексты научного жанра, а именно – ВКР студентов одного из вузов РФ. Лексическая плотность, по утверждению авторов, проявляет низкую корреляцию ($k \leq 0.3$) с лексической насыщенностью: другими словами, лексическая плотность почти не зависит от употребляемых лексических единиц. Аналогичные параметры исследуются на материале художественных текстов в работе [Орехов, 2022]. Данная идея проверяется и для текстов интернет-пространства, тем более, для них характерно отсутствие строгих жанровых и языковых рамок для создания текстов.

В качестве инструмента для сбора количественных показателей использовалось приложение Profiling-UD. Расчеты были проведены в среде Excel.

1. В начале необходимо определить, принадлежат ли количественные параметры выборки нормальному распределению. Для этого воспользуемся критерием Колмогорова-Смирнова [Lilliefors, 1967] с учетом дисперсии по формуле (5).

$$D_n^* = D_n(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}) \quad (5)$$

В формуле (5) D_n– дисперсия исследуемой выборки значений, n – количество элементов в выборке, D_n^* – экспериментальное значение критерия, которое сравнивается с критическим.

2. Если распределение ненормальное, то для расчета силы корреляции используется коэффициент ранговой корреляции Спирмена.

3. Если распределение нормальное, то для расчета силы корреляции используется коэффициент корреляции Пирсона.

Конечно, необходимо учитывать, что могут существовать и факторы-посредники, влияющие на корреляцию, как заявляют авторы исследования [Köhn, Baumann, Dörfler, 2018]. Они провели эксперименты по выявлению корреляций синтаксических и просодических параметров. Отмечено, что потенциальная разница в высоте голоса у подлежащих и дополнений может возникнуть из-за того, что индоевропейские языки, как правило, имеют порядок SVO, т.е. дополнения появляются позже подлежащих, а высота голоса имеет тенденцию к уменьшению к концу повествовательного предложения (так называемый low fall в англоязычной традиции, ИК-1 в классификации Е.А. Брызгуновой). В настоящем исследовании представлены общие взаимосвязи.

Мы рассчитали корреляции для 23 скрытых сообществ из 34, что связано с малым количество участников (4 и менее): «Астрономия», «Биология», «География», «Журналистика», «Информатика», «Машиностроение», «Производство», «Социология», «Техника», «Филология», «Философия». В разделе 3.8 исследования представлены итоговые профили сообществ. В случае невозможности проведения корреляционного анализа значения не присваиваются ни одному из исследуемых параметров. Если в ходе расчетов уровень значимости $p > 0.05$, то корреляция считалась незначимой, а вместо значения ставился прочерк.

3.7 Лингвистические профили скрытых сообществ

3.7.1 Скрытое сообщество «Армия и государственная безопасность»

В данном сообществе для расчета корреляций использовались выборки из постов 17 пользователей, которые являются членами данного скрытого сообщества. Все семь

пар выборок в результате подсчетов подчиняются ненормальному распределению, из них пять пар имеют значимые корреляции при уровне значимости $p < 0.05$. На диаграммах с ненормальным распределением в качестве значений представлены ранги. Сводные количественные данные по всем скрытым сообществам и проанализированным корреляциям представлены в таблице 6 раздела 3.8 диссертации, в следующих разделах будут даны лингвистические комментарии полученным результатам.

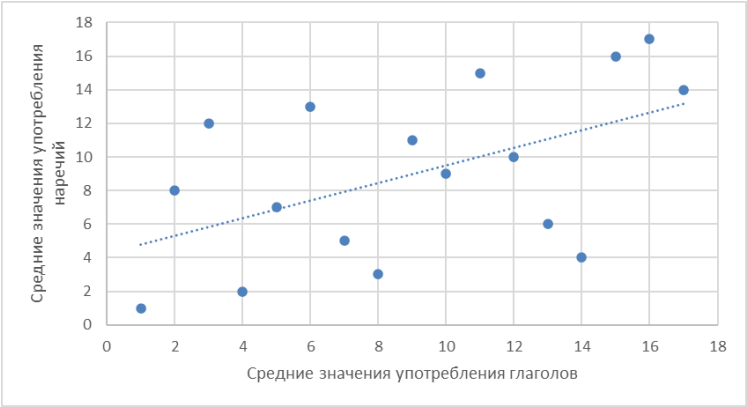

Рисунок 15 — Внутритекстовая корреляция средних значений употреблений глаголов и наречий

На рисунке 15 точками представлены ранговые показатели усредненных значений употребления глаголов и наречий, а линия тренда используется как дополнительный статистический показатель коэффициента корреляции. Далее в исследовании визуализированные корреляционные структуры будут представлены аналогичным образом. Средние значения употреблений глаголов и наречий обладают умеренной связью с коэффициентом $r = 0.5245$. В текстах пользователей действительно наблюдаются тенденции к усилению/ослаблению значения действий с помощью таких наречий, как «немного», «часто», «максимально» и др. Приведем тематические примеры для пользователей с ID 291669 и 55040. Все примеры здесь и далее по тексту исследования приводятся в авторской орфографии и пунктуации.

1. ID 291669: *«…Мы с вами - простой народ. Мы* ***всегда*** *будем под ударом. Умирают люди. ПОКА ЧТО - это действительно военная операция. Не плюйтесь, это действительно так. Но,* ***ещё немного*** *и это может перерасти в реальную войну со всеми вытекающими. В этой связи не радует, пугает то, что*

стороны ***пока*** *не могут сесть за стол переговоров, хотя якобы пытаются. Надеюсь, что в последних трёх постах я довёл свою точку зрения на данный вопрос. Пояснять* ***больше бы*** *не хотелось. Моя позиция* ***максимально*** *чуткая и гибкая, но не стоит путать её с прогибающейся, уж очень* ***часто*** *мрази путают доброту и слабость…»*

2. ID 55040: *«…****Сегодня вновь*** *выезжал в город.* ***Понемножку*** *очищают улицы ото льда и снега, расчищают трамвайные пути. Работают исключительно женщины. Вообще "бабья" рука за период войны и многое сделала, и многое спасла... Если по соседству есть холостяки, и они думают все же обзаводиться когда-нибудь семьей, женой, – советуйте им всем: пусть женятся на ленинградских девушках…»*

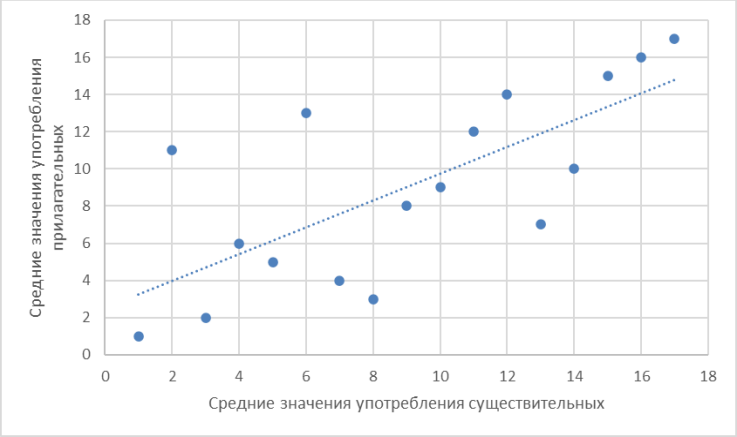

Рисунок 16 — Внутритекстовая корреляция средних значений употреблений имен существительных и имен прилагательных

Для пары «существительное-прилагательное» наблюдается более сильная положительная корреляция $r = 0.7205$ (рис. 16). Прилагательные-интенсификаторы наблюдаются в большом количестве не только в вышеприведенных примерах, но и у других пользователей скрытого сообщества. Например, у пользователя с ID 339663: *«…Я не знаю, что докладывает МО РФ Верховному Главнокомандующему, но по моему* ***личному*** *мнению, надо принимать более* ***кардинальные*** *меры, вплоть до объявления* ***военного*** *положения на приграничных территориях и использования* ***маломощного ядерного*** *вооружения…»*

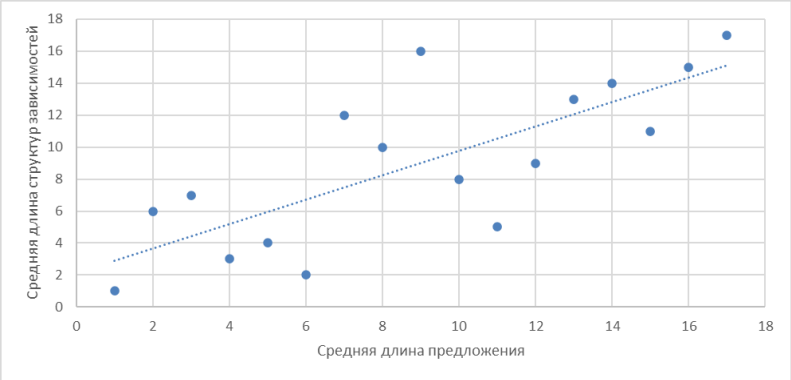

Рисунок 17 — Внутритекстовая корреляция средних значений длины предложений и длины структур зависимостей

С точки зрения синтаксиса наблюдается увеличение длины предложения, что влечет за собой увеличение средней длины структур зависимостей (в терминологии Г.Я. Мартыненко и А.О. Гребенникова – *‘степени дистантизации’* [Мартыненко, Гребенников, 2018, с. 20], рис. 17), в частности, при препозитивных атрибутивных группах. Линейную степень дистантизации можно проецировать на глубину дерева от определенного узла в иерархической структуре, полученной в формате CoNLL-U – одного из форматов отображения морфосинтаксической структуры предложения. Так, для предложения из поста пользователя с ID 2970 степень дистантизации между узлами «на» и «корабли» равняется трем, что соответствует трем узлам дерева зависимостей, стоящими между ними (рис. 18).

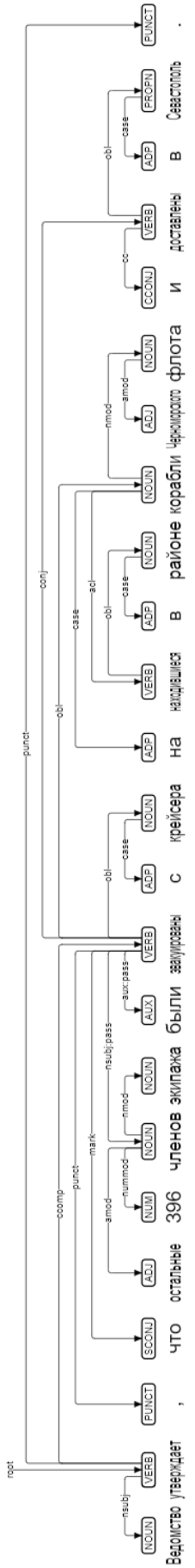

Рисунок 18 — Дерево зависимостей для предложения из поста пользователя с ID 2970

В данном подкорпусе не выявлена лексическая корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия». При $r = -0.3395$ был вычислен $p = 0.18$, что указывает на недостаточное количество данных для подтверждения значимости корреляции.

3.7.2 Скрытое сообщество «Бизнес, коммерция, экономика, финансы»

В данном сообществе 52 пользователя, которых можно охарактеризовать по пяти значимым корреляциям из семи при уровне значимости $p < 0.05$. Для имен существительных и глаголов наблюдается обратная взаимосвязь средней силы: при росте употреблений имен существительных снижается число употреблений глаголов (рис. 19).

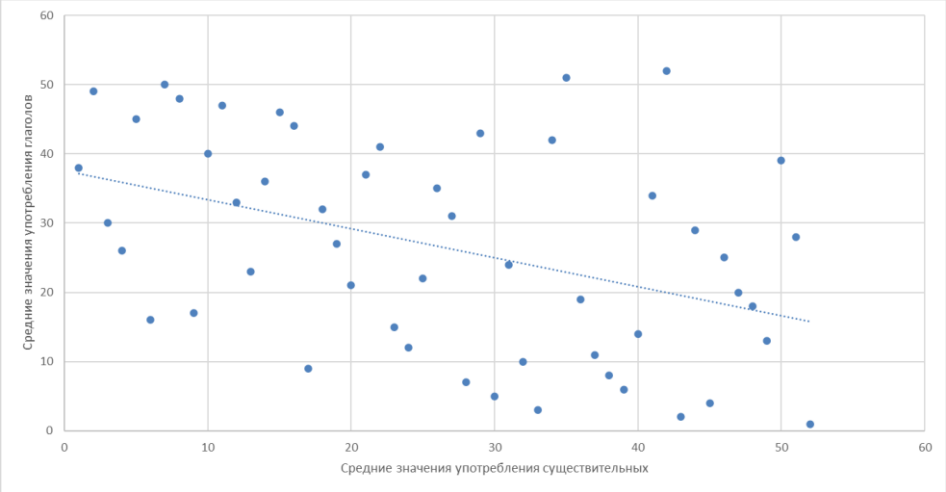

Рисунок 19 — Внутритекстовая корреляция средних значений употреблений имен существительных и глаголов

Подобные случаи связаны со способом построения постов, которые пользователь описывает в номинативных конструкциях, например, предстоящие мероприятия (см. пример постов пользователей с ID 13977 и 20518 соответственно). В первом примере с помощью существительных в рамках составного именного сказуемого перечисляются все профессии и заслуги личности, а во втором – в виде маркированного списка перечисляются идеи, которые предлагает пользователь. Несмотря на то, что во втором примере встречаются два глагола («даю», «впишется»), количество имен существительных значительно выше.

1. ID 13977: *«…15.50-16.10 «ЛИЧНЫЙ БРЕНД В ПОМОГАЮЩИХ ПРОФЕССИЯХ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ»*

*Солдатова Светлана (Симферополь) - **бизнес-консультант**, **коуч**, управляющий **партнёр** консалтинговой компании «S&D Group", действительный **член** ППЛ…»*

2. ID 20518: *«…Вот **парочка** идей, я даю очень много пользы и возможностей безвозмездно!*

*•Дополнительный **доход**, высокая маржинальность.*

•Отлично впишется в интерьер празднично лофта, ресторана или островка в ТЦ.

*•Дополнительная **фотозона** при проведении праздников.*

*•Дополнительная развлекательная **программа** для детей.*

*•Дополнительная **витрина** для размещения товаров (игрушки, воздушные шарики, сладости)…»*

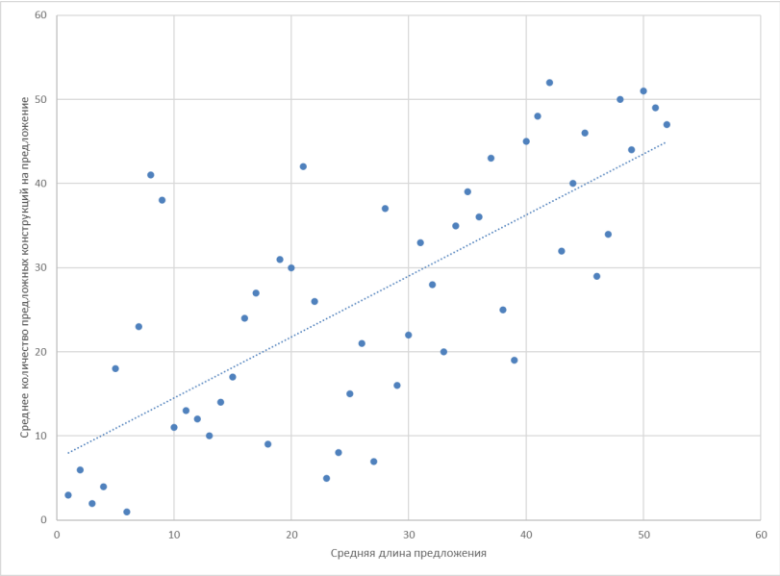

Рисунок 20 — Внутритекстовая корреляция средних длин предложений и среднего количества предложных конструкций

На синтаксическом уровне наблюдается сильная положительная корреляция длины предложения и среднего количества используемых предложных конструкций (рис. 20). Необходимо отметить, что на эту корреляцию может повлиять и длина предложных конструкций, что подтверждено исследованиями разножанровых текстовых коллекций на материале русского и английского языков [Хохлова, Рубинер, 2019; Curtotti, McCreath, 2011, p. 205]. В постах данного скрытого сообщества также встретились длинные предложные конструкции, на

размер которых повлияло употребление нескольких атрибутивов перед зависимым существительным.

1. Длиной в пять единиц (пользователь с ID 32764): *«…А потом **в одно не прекрасное утро** понимает, что «всё не то». Ничто не изменилось, но он перестал быть счастливым…»*

2. Длиной в шесть единиц (например, пользователь с ID 32764): *«…**В любой непонятной или спорной ситуации** он будет замирать и ждать того, кому видней. Внимание, вопрос: разве сможет такой человек быть эффективным сотрудником, особенно в 21 веке, где постоянно нужно принимать решения и перепроверять информацию?...»*

Тем не менее, в подкорпусе встречались и сложные предложения с большим количеством коротких предложных конструкций (см. пост пользователя с ID 243192): *«…ВАЖНО: оповещайте **об отмене** сеанса **за 1-2 суток**, ситуации **у всех** бывают разные, поэтому сообщайте, пожалуйста, **обо всех изменениях** заранее…»*.

Лексическая корреляция, как и в случае с сообществом про армии, не была выявлена ($r = 0.1455$, $p = 0.3$).

3.7.3 Скрытое сообщество «Дом и домашнее хозяйство»

Для постов 35 пользователей выявлено пять значимых корреляций при $p < 0.05$. На обобщающем рисунке 21 видно, что морфологические корреляции (второй и третий ряды) обладают умеренной связью: $r = 0.4661$ для существительных и прилагательных и $r = 0.4008$ для глаголов и наречий, при этом встречается и отрицательная корреляция существительных и глаголов приблизительно с той же силой связи – $r = -0.4784$. Например, пользователь с ID 85999 приводит описания различных рецептов, поэтому в начальных частях текста при описании ингредиентов имена существительные будут преобладать над глаголами, при этом список ингредиентов может быть как немаркированным, так и маркированным. Пример публикации приведен после рисунка 21.

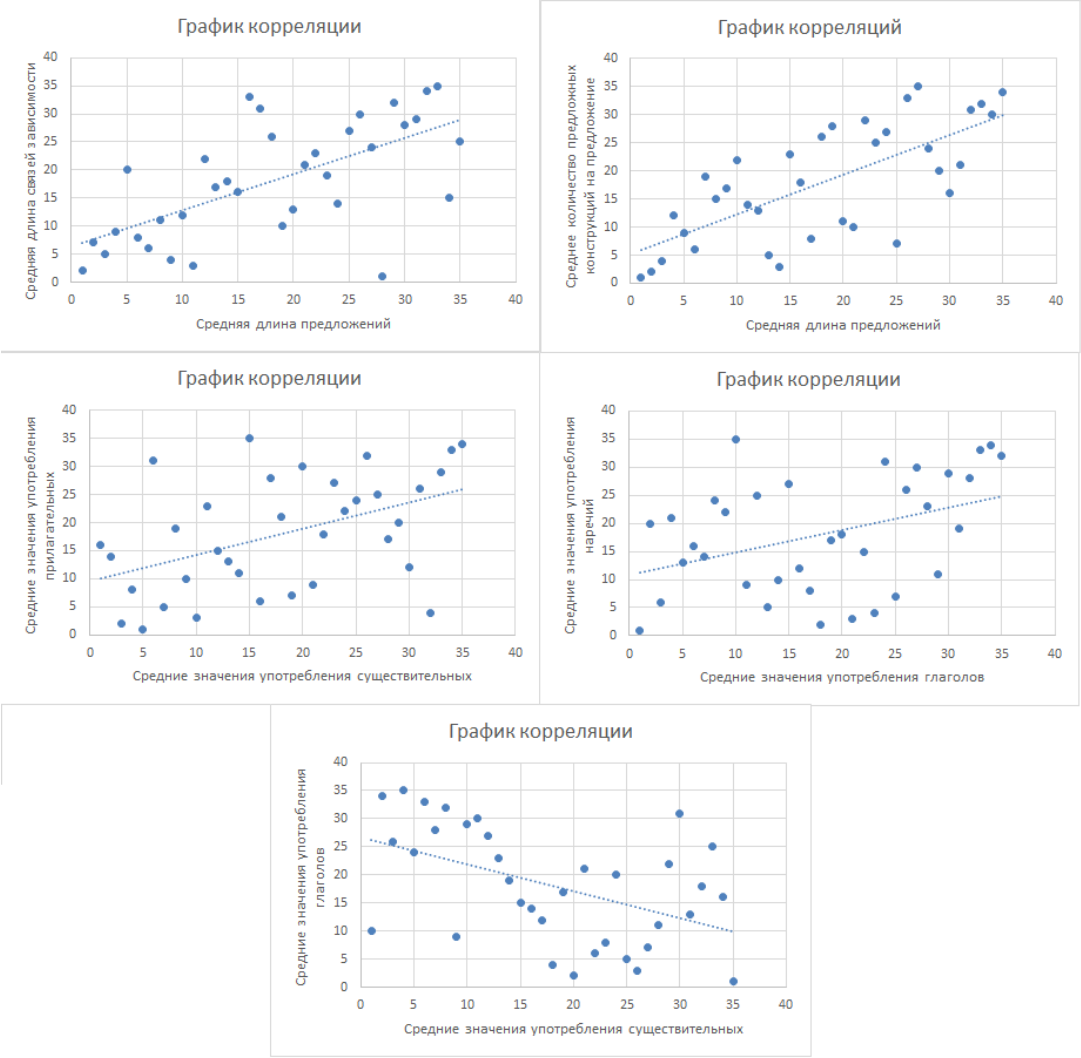

Рисунок 21 — Сводные данные о внутритекстовых корреляциях

1. ID 85999: *«…**Рецепт форшмака**. 1 **сельдь**, 1 варёная **морковка**, 1 **луковица**, 100 **грамм масла** сливочного, 1 **яйцо**, **кусочек** белого **хлеба**, размоченного в воде взбиваем в блендере или через мясорубку и получаем прекрасный форшмак. Приятного аппетита…»*

2. ID 85999: *«…**Рецепт Том-яма***

***Ингредиенты**:*

*- **бульон** куриный 2,5 литра*

*- кокосовое **молоко** 400 мл*

*- рыбный **соус** 6 ст.**ложек***

*- **лемонграсс** 6 **стручков***

*- **цедра** одного **лайма***

*- **сок** из 2 **лаймов***

- ***вешенки*** *300 гр*
- ***шампиньоны*** *300 гр*
- *4* ***листика каффир-лайма***
- ***корень галанга***
- ***корень имбиря***
- ***сахар*** *1 ст.****ложка***
- *2* ***перца чили***
- ***креветки*** *королевские очищенные 500 гр*
- ***паста*** *для том-яма 100 гр…»*

Для синтаксических параметров характерна положительная корреляция выше 0.6: для длины предложения и степени дистантизации – 0.6414, для длины предложения и количества предложных конструкций – 0.7081. Несмотря на самое высокое значение корреляции в данном подкорпусе, существуют и отклонения: даже для сравнительно коротких предложений (например, в 13 словоформ) наблюдается большое количество предложных конструкций: *«…Маленький магазинчик притаился* ***на въезде в Новосаратовку***, ***вдали от домов с шаговой доступностью****…»* (предложение из поста пользователя с ID 785629). Тем не менее такие случаи единичны и не оказывают влияния на общую синтаксическую тенденцию.

Как и в прошлых сообществах, в лингвистическом профиле отсутствует значимая лексическая корреляция.

3.7.4 Скрытое сообщество «Досуг, зрелища и развлечения»

В этом сообществе были взяты выборки из постов 169 пользователей, которые стали членами данного скрытого сообщества, для вычисления корреляций, оно является наиболее крупным сообществом в исследовании. После проведения вычислений все семь пар выборок также распределены ненормально, из них пять пар выборок обладают статистически значимыми корреляциями при $p < 0.05$ (см. сводные данные на рисунке 22). В отличие от предыдущих сообществ, все корреляции являются положительными. Так, в сообществе «Армия и

государственная безопасность» внутритекстовая корреляция между именами прилагательными и наречиями была отрицательной с умеренной связью и равнялась $r = -0.5196$. При увеличении именных модификаторов уменьшалось количество глагольных модификаторов, например: *«…В армии нужно назначать командирами людей* ***твердого*** *характера,* ***смелых***, ***принципиальных***, *которые переживают за своих бойцов, которые зубами рвут за своего солдата, которые знают, что подчиненного нельзя оставлять без помощи и поддержки…»* (предложение из поста пользователя с ID 339663). Однако в скрытом сообществе любителей развлечений наблюдается слабая тенденция к взаимному росту как имен прилагательных, так и наречий ($r = 0.1742$): *«…Воскресение. Время обеда. Идём потихоньку к машине наслаждаясь водой и видами* ***сайменского*** *канала. Ночевали на острове где много* ***мелких*** *комариков и стрекоз…»*

Рисунок 22 — Сводные данные о внутритекстовых корреляциях

Оставшиеся четыре пары показали положительные корреляции с умеренной связью в диапазоне $r \in [0.4044; 0.6342]$, самое большое значение пришлось на пару «длина предложения-количество предложных конструкций».

1. Пользователь с ID 42744: *«…**В рамках** дневной программы Dance Exchange New Stage состоится преселекшн (отбор) **на баттл** и **в этом отборе** могут принять участие танцоры любых стилей, **независимо от пола, возраста, танцевального опыта** и тд…»* (четыре предложные конструкции на 30 словоформ).

2. Пользователь с ID 39065: *«…Активный отдых **на природе** это отличная возможность забыть **о повседневных проблемах**, отойти **от суеты**, перезагрузиться и **с пользой для здоровья** провести время!…»* (пять предложных конструкций на 22 словоформы).

3. Пользователь с ID 21530: *«…**За одним столом** семья **из 13-ти человек**, Снегурочка **в роли** меня **с разными заданиями для детей** и взрослых, а так же, ставший уже традицией **для детей - за час до боя** Курант дети проходят Квест **в квартире по поиску** клада, от того веселее, нет утомительного ожидания, а **в конце** их всегда ждёт приз…»* (11 предложных конструкций на 52 словоформы).

Лексическая корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия» оказалась незначимой, корреляционные связи между переменными отсутствуют: $r = 0.0718, p = 0.3534$.

3.7.5 Скрытое сообщество «Здоровье и медицина»

Данное сообщество с точки зрения количественных параметров оказалось одним из наиболее полных (наравне со скрытыми сообществами про образование, право и спорт), так как, во-первых, оно представлено шестью значимыми корреляциями из семи возможных, во-вторых, в нем представлена лексическая корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия», которая присутствует еще в трех других сообществах: «Образование», «Спорт» и «Религия». Сводные данные представлены на рисунке 23.

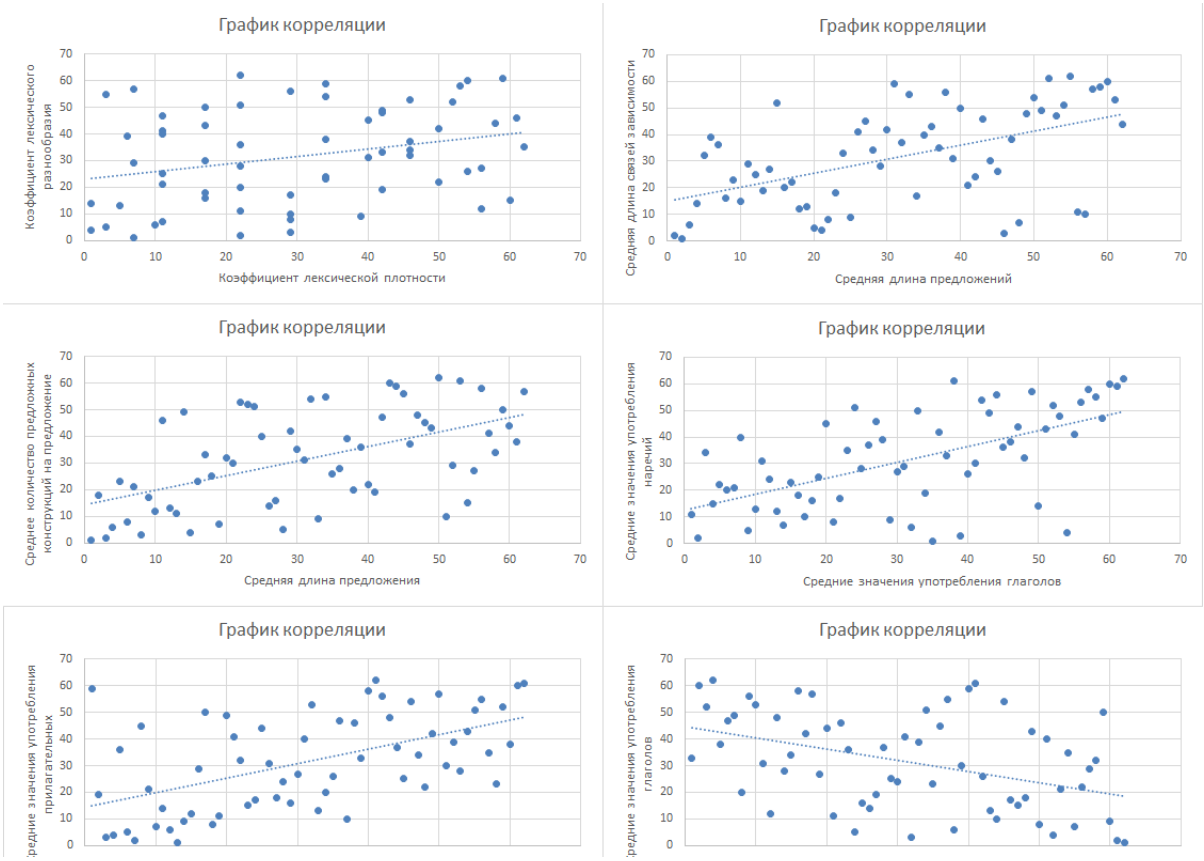

Рисунок 23 — Сводные данные о внутритекстовых корреляциях

Положительная корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия» (первый ряд, левая колонка, $r = 0.286, p = 0.0241$) указывает на возможность богатого лексического наполнения текстов, создаваемых пользователями данного сообщества, а также может служить дополнительной характеристикой при оценке коммуникативно-письменных компетенций пользователя: например, для оценки навыков индивидуального выражения идей и мыслей [Николенкова, 2007; Каинова, 2018]. Для пользователя с ID 519589 коэффициент лексической плотности равняется 0.67, а коэффициент лексического разнообразия равен 0.9, что может свидетельствовать об информативности его постов и богатстве словаря. В следующем отрывке повествуется о личном опыте борьбы с COVID-19, наблюдается использование междометий («мол») наравне со знаменательными и служебными частями речи, а также использование жаргонизмов («косарь»): *«…В один момент не выдержала и заказала ускоренный платный ПЦР за косарь и привет. Зато моментально позвонили из откуда следует, сказали, что по телефону будут отслеживать*

перемещения, посулили врача и пожелали доброго здоровьичка. Врач через полчаса перезвонил, удивился, что с голосом капец - мол на практике такого не встречал - послал лучи добра и Яндекс Курьера с бесплатными лекарствами. Предвещая поток статей и комментарии про средство - знаю, вижу побочки, да и врач попросил только в крайнем случае прикасаться к этому. И к парацетамолу тоже…»

Незначимой отрицательной корреляцией обладают пары значений имен прилагательных и наречий ($r = -0.2082, p = 0.1042$), значимой отрицательной корреляций обладают пары значений имен существительных и глаголов ($r = -0.4228, p = 0.0006$), это значение сопоставимо с ранее выделенными корреляциями для других сообществ: $r = -0.4182$ для сообщества «Бизнес, коммерция, экономика, финансы» и $r = -0.4784$ для сообщества «Дом и домашнее хозяйство». Оставшиеся корреляции сообщества про здоровье положительны и обладают приблизительно одинаковыми силами связи: $r = 0.5434$ для пары «имя существительное-имя прилагательное», $r = 0.5979$ для пары «глагол-наречие», $r = 0.5239$ для пары «длина предложения-степень дистантизации» и $r = 0.5481$ для пары «длина предложения-количество предложных конструкций».

3.7.6 Скрытое сообщество «Искусство и культура»

В сообществе, в которое входят 115 пользователей, выявлено три значимые морфологические корреляции и две значимые синтаксические корреляции (рис. 24).

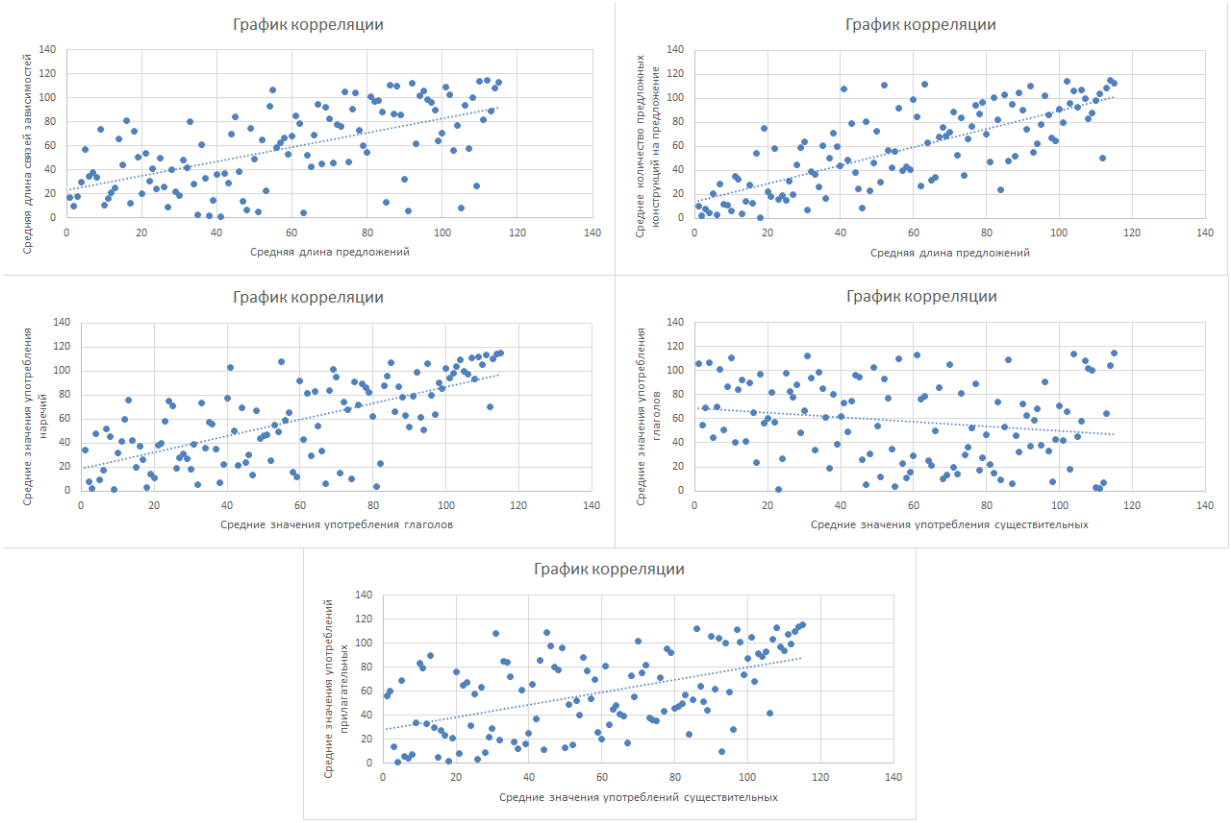

Рисунок 24 — Сводные данные о внутритекстовых корреляциях

Среди них лидирующую позицию занимает связь длины предложения с количеством предложных конструкций (первый ряд, правая колонка): $r = 0.7565$.

1. Пользователь с ID 871024: *«…Vasya Ve - Fiesta the Pogues - я впервые* ***за много лет*** *услышал песню, которая написана не* ***для чартов*** *и чтобы кого-то удивить или высказаться* ***о боли****, а просто весёлый балаган,* ***в котором*** *творчество* ***ради творчества****…»* (5 предложных конструкций на 34 словоформы).

2. Пользователь с ID 68068: *«…Александр Лециус и Кристина Карпышева* ***за последние несколько лет*** *показали свои знаменитые аудио-визуальные перформансы* ***на Таймс-сквер, в Музее искусств Лос-Анджелеса, на фестивале медиа искусств в Японии, на монреальском фестивале MUTEK*** *— Каннах* ***в области цифрового искусства и электронной музыки****…»* (7 предложных конструкций на 38 словоформ).

3. Пользователь с ID 22242: *«…Кульминация* ***в сценарии****, должна быть зримой* ***на сцене****, она не должна быть словами* ***в тексте ведущего****…»* (3 предложные конструкции на 16 словоформ).

Положительная сила взаимосвязи также выявлена у глаголов и наречий-модификаторов (второй ряд, левая колонка): $r = 0.6791$. Вслед за ним с $r = 0.5946$ установлена корреляция длины предложения и степени дистантизации (первый ряд, левая колонка). Так, для предложения длиной в 18 словоформ из поста пользователя с ID 69784 степень дистантизации между главным узлом «из» и подчиненным узлом «звука» равна 4, при этом в иерархической структуре предложения (рис. 25) за счет сочинительных связей между исследуемыми узлами наблюдается несколько ветвей составляющих, в отличие от того, что ранее было представлено на рисунке 18.

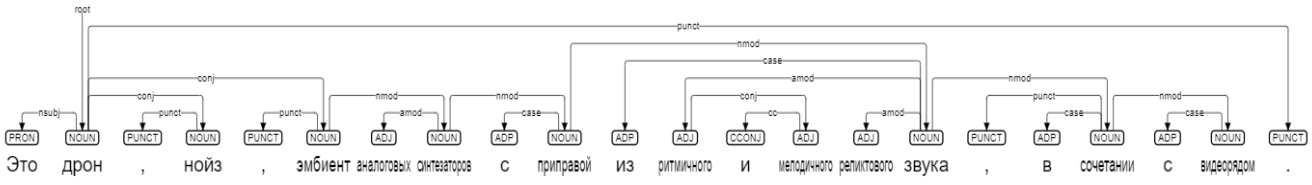

Рисунок 25 — Дерево зависимостей для предложения из поста пользователя с ID 69784

Одна из отрицательных корреляций – взаимосвязь существительных и глаголов (второй ряд, правая колонка), однако в этом сообществе не наблюдается четких тенденций из-за слабой связи ($r = -0.1924$).

3.7.7 Скрытое сообщество «История»

В тематическом сообществе из 23 пользователей выявлено четыре значимых корреляции (рис. 26).

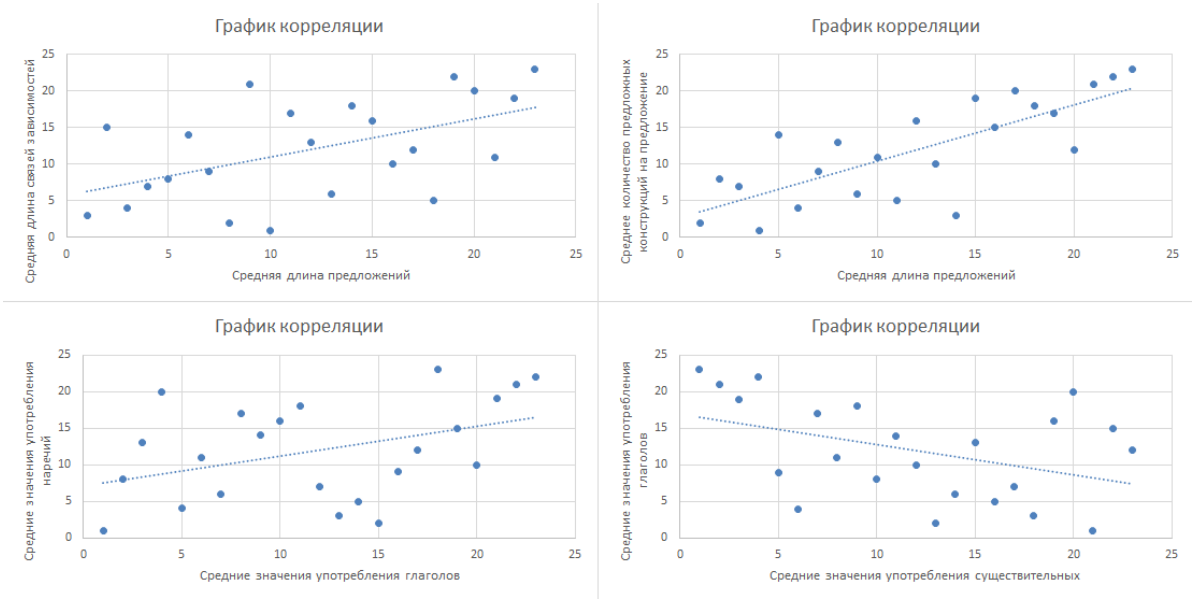

Рисунок 26 — Сводные данные о внутритекстовых корреляциях

Для взаимосвязи существительных и глаголов выявлен отрицательный коэффициент с силой связи $r = -0.416$ (второй ряд, правая колонка). Например, в посте пользователя с ID 8436 в приведенном предложении на 11 существительных приходится три глагола, что связано не со способом оформления текста (ранее была отмечена одна из причин выявления таких корреляций – подача информации в виде маркированного текста), а с наличием предложно-именных групп и дополнений: *«…Так же **книги** про **оружие** и **пулеметы** тех **годов** когда официальная **история** нам преподносит **войну** с **Наполеоном**, когда они с голой **жопой** перекатывали **пушки** на деревянных **лафетах**, которые стреляли **ядрами**…»*.

Для глаголов и наречий (второй ряд, левая колонка) – $r = 0.4041$, для длины предложений и степени дистантизации (первый ряд, левая колонка) – $r = 0.5197$, для длины предложений и количества предложных конструкций (первый ряд, правая колонка) – $r = 0.7717$. Для корреляции имен существительных и прилагательных выявлен пограничный уровень значимости $p = 0.6$, в результате чего она не включена в итоговый профиль.

3.7.8 Скрытое сообщество «Легкая и пищевая промышленность»

На рисунке 27 приведены корреляции исследуемых групп признаков. Согласно подсчетам, было выявлено пять статистически значимых корреляций, три из которых относятся к положительным, две — к отрицательным. Коэффициент большинства корреляций равен или выше 0.7, в том числе и отрицательный коэффициент для группы имен существительных и глаголов (второй ряд, правая колонка) — $r = 0.7069$, т.е. наблюдается тенденция к сильной связи. В группе имен существительных и имен прилагательных коэффициент корреляции самый высокий из всех представленных в группе (второй ряд, левая колонка) — $r = 0.8473$.

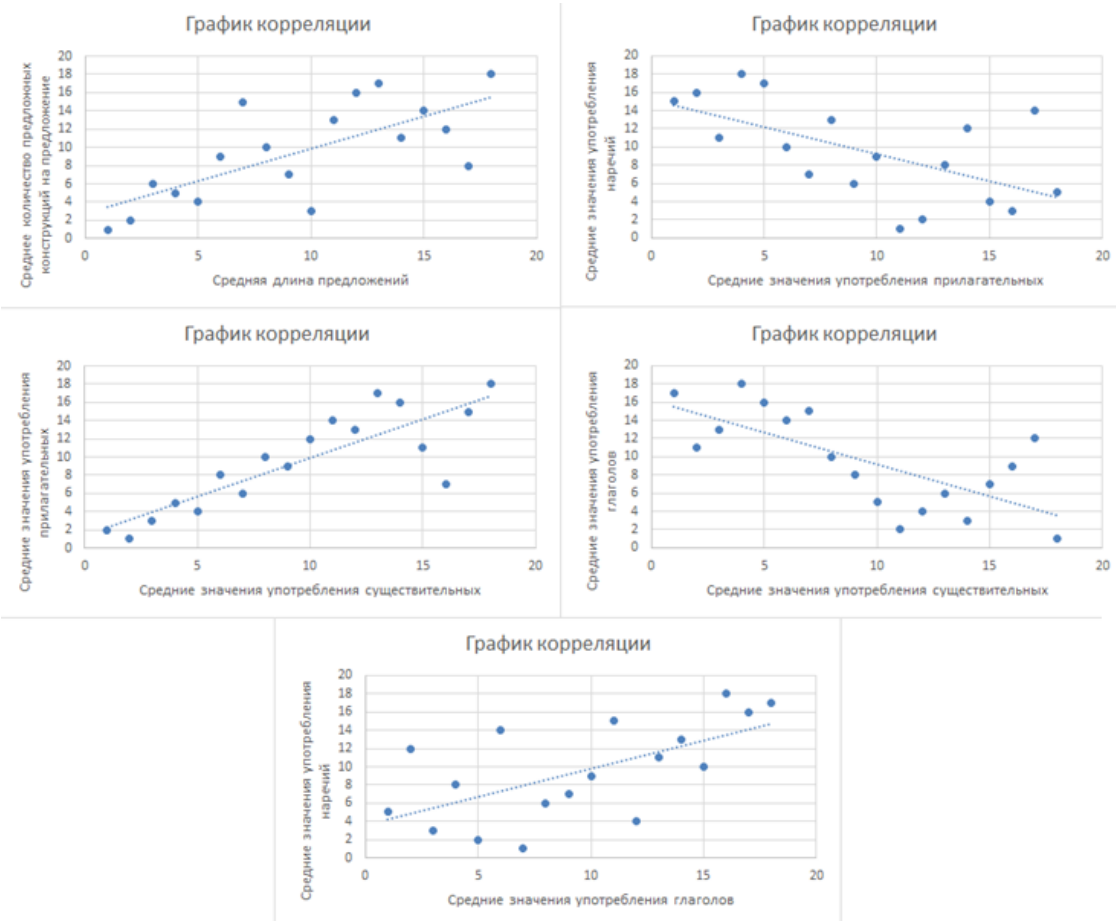

Рисунок 27 — Сводные данные о внутритекстовых корреляциях

В частности, некоторые посты пользователя с ID 2303 являются рекламными. Для подобного типа текстов, согласно [Волобуев, 2013], характерно использование различных форм имен прилагательных для усиления положительных коннотаций, привлечения целевой аудитории, а также и создания уникального образа продукта или группы продуктов: *«…Мы снова будем рады вас видеть в нашем кафе!!! Где вас ждёт* ***приятная*** *атмосфера,* ***горячий*** *кофе,* ***любимые*** *вафли,* ***сытные*** *сэндвичи и* ***освежающие*** *лимонады…»*

Такие же рекламные посты встречаются у пользователя с ID 30766: *«…Весь мёд упакован и расфасован в* ***небольшие*** *баночки, как и в* ***прошлом*** *году. И это может стать* ***прекрасным милым*** *подарком в* ***новый*** *год для близких и друзей. А может и просто* ***вкусным*** *дополнением к чаю, или* ***лечебной*** *ложечкой утром натощак…»*. Для этого же пользователя характерно использование в ряде постов двусоставных предложений с составным именным сказуемым, в результате чего снижается количество глаголов, которые используются в тексте: *«…В этом году у*

нас есть некоторое количество ёлочек для Вас! ***Высота 54 см Ширина 34 см Глубина 10 см****…».*

Самый низкий коэффициент корреляции встречается в группе прилагательных и наречий (первый ряд, правая колонка) — $r = -0.5933$, что указывает на обратную взаимосвязь. В постах о текстильной промышленности пользователя с ID 35300 наречия-модификаторы встречаются в простых распространенных предложениях, а прилагательные-модификаторы – в сложных предложениях, при этом характерны случаи, когда прилагательные или наречия могут попросту не употребляться в предложениях, например: *«…С механизацией* ***текстильной*** *промышленности* ***производственные*** *затраты на одежду падали, и поэтому возник фаст-фэшн…», «…В результате одежду быстро меняли или выбрасывали…», «…И количество секонд-хенд магазинов очень быстро росло.…».*

3.7.9 Скрытое сообщество «Наука и технологии»

В корпусе всего 12 пользователей публиковали посты, связанные с научной тематикой (рис. 28).

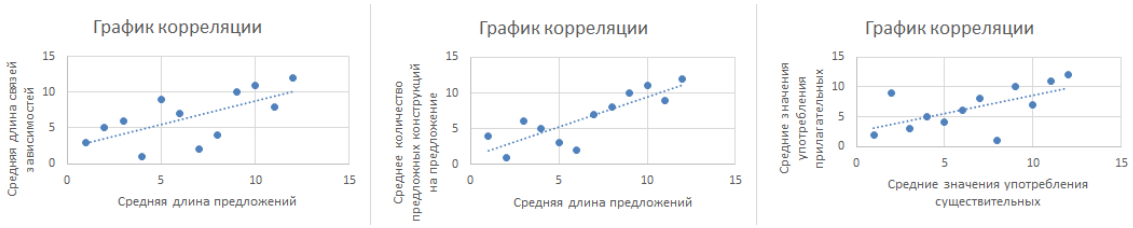

Рисунок 28 — Сводные данные о внутритекстовых корреляциях

Этот подкорпус является одним из наименее детализированных, поскольку было выделено всего три значимые статистические корреляции для ненормально распределенных выборок: группа «имена существительные-имена прилагательные» (правая колонка, $r = 0.6084, p = 0{,}036$), группа «длина предложения-степень дистантизации» (левая колонка, $r = 0.6503, p = 0{,}022$) и группа длина предложения-количество предложных конструкций» (центральная колонка, $r = 0.8392, p = 0{,}0006$). Например, часть постов пользователя с ID 27 посвящена развитию законов диалектики и их практическому применению. Увеличение степени дистантизации между главным и зависимым узлами в постах его профиля связано с наличием препозитивных определений, выраженных

именами прилагательными, перед дополнениями. Например, степень дистантизации для приведенного на рисунке 29 предложения между узлами «с» и «учеными» равна трем при общей длине предложения в девять узлов синтаксического дерева. Факт наличия прилагательных-модификаторов также влияет и на положительную корреляцию с именами существительными. Пользователь с ID 8436 на своей странице кратко повествует о содержании книги по современным технологиям, при этом название каждой приводимой им отрасли снабжено прилагательным, которое является неотъемлемой частью названия, например: *«**поотраслевая** (sic!) логистика»*, *«теория **размерной** (sic!) обработки»*, *«патенты (sic!) и **патентное** право (sic!)»* и пр.

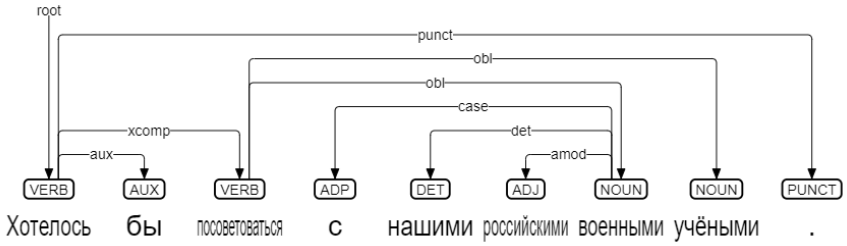

Рисунок 29 — Дерево для предложения из поста пользователя с ID 27

3.7.10 Скрытое сообщество «Образование»

При анализе данного сообщества, в котором «состоят» 73 участника, выявлено шесть значимых групп признаков на морфосинтаксическом и лексическом уровнях, что подробно представлено на рисунке 30. В отличие от коэффициентов, полученных для сообщества «Здоровье и медицина», значения в этом сообществе находятся в диапазоне $r \in [-0.3633;\ 0.5233]$: они ниже, что свидетельствует о слабой связи или ее отсутствии. Правая граница диапозона является коэффициентом для группы «длина предложения-количество предложных конструкций», левая граница – для группы «имя существительное-глагол». Коэффициент корреляции для пары «коэффициент лексической плотности-коэффициент лексического разнообразия» на несколько единиц ниже ($r = 0.2262$), чем в случае для сообщества «Здоровье и медицина» ($r = 0.286$), но тем не менее, вводится предположение, что посты пользователей этого сообщества будут информативнее и богаче по лексическому составу, чем посты пользователей

других сообществ, в которых данная корреляция была незначимой. Отчасти такое поведение можно объяснить типом созданного текста – рекламно-образовательный, целью которого является привлечение потенциальных учеников в свои организации, например, пост пользователя с ID 562755: *«…Тебе от 13 до 18 лет, очень хочешь познать, что такое диджеинг, мечтаешь покорить танцполы, удивить своих друзей, выступать на различных площадках или ты просто неравнодушен к музыке и различным крутилкам, то тогда тебе к нам… Вы узнаете*

- классификация музыкальных стилей;

- сведение композиций на виниле и cd;

- теория построения миксов;

- PR и реклама себя как артиста... »

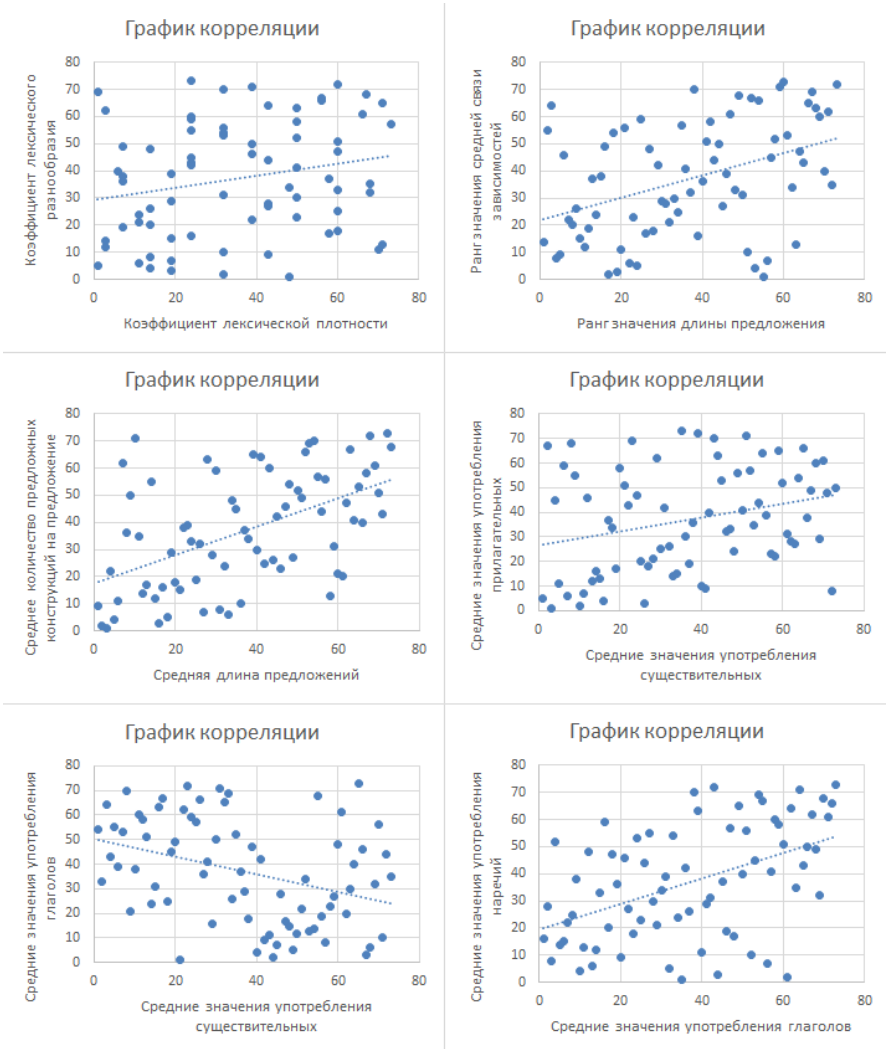

Рисунок 30 — Сводные данные о внутритекстовых корреляциях

3.7.11 Скрытое сообщество «Политика и общественная жизнь»

В этом сообществе состоит 33 участника, для чьих постов было выделено четыре значимые корреляции из семи. На рисунке 31 представлена информация, показывающая, что морфологические взаимосвязи имеют умеренную степень зависимости: коэффициент корреляции $r = 0.4856$, уровень значимости $p = 0.004$ для существительных и прилагательных, а также $r = 0.6715$, уровень значимости $p = 0{,}0000189$ для глаголов и наречий. Как и во многих других сообществах, наивысший положительный коэффициент корреляции наблюдается в случае пары «длина предложения-количество предложных конструкций»: $r = 0.7179$, $p = 0{,}00000256$. Такие показатели объясняются одной из подтематик-доминантой сообщества, а именно – российском-украинскому конфликту. Свои настроения пользователи передают через использование эмоционально окрашенной лексики (за счет прилагательных и наречий), например: *«…Сегодня я убедился в своей правоте - Украина как государство погибло в 2014 году, а на смену ему пришёл* ***нацистский*** *режим, который взял в заложники население этой* ***замечательной*** *страны, в которой я ни раз до этого бывал...»* (пост пользователя с ID 3450), *«…Вы замечали, что лжецы, со временем глупеют? Тому пример* ***хитроумные*** *планы Запада по захвату России. Чем* ***активней*** *они лгут, тем* ***глупее*** *и* ***примитивные*** *становятся их же население…»*

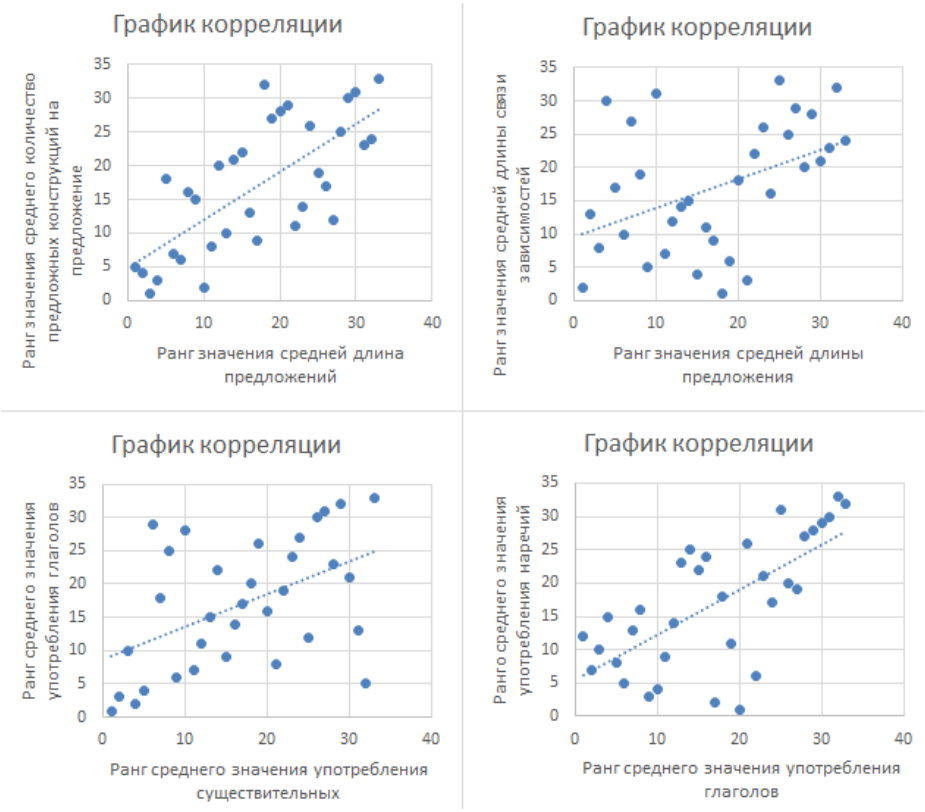

Рисунок 31 — Сводные данные о внутритекстовых корреляциях

3.7.12 Скрытое сообщество «Право»

Данное сообщество представлено 12 пользователями, которые публикуют посты на такие подтематики, как правовые статусы объектов, работа в органах госслужбы, правонарушения и пр. Лингвистический профиль представлен шестью корреляциями из семи, исключением является группа «коэффициент лексической плотности-коэффициент лексического разнообразия». Синтаксические коэффициенты корреляции достаточно высоки. Например, для группы «длина предложения-количество предложных конструкций» значение $r = 0.8182$, $p = 0{,}0012$. Данную особенность можно объяснить тем, что пользователи в собственные посты могут добавлять отрывки из правовых текстов, в частности, из разных кодексов РФ. Современные правовые тексты имеют более сложную синтаксическую структуру, чем в начале их создания в современной России [Савельев, 2020], что приводит к синтаксическому усложнению пользовательских постов. Так, в приведенном отрывке пользователь средствами косвенной речи обсуждает основные пункты уголовного кодекса. Наличие жаргонных лексических единиц указывает на то, что ссылка на кодекс закончилась и начались собственные

мысли пользователя: *«…Согласно статье 301 УК РФ, сотрудник правоохранительных органов, задержавший гражданина, не имея на то достаточных оснований, может быть наказан лишением свободы на срок до двух лет. А вот, если человека незаконно заперли в* ***кутузку*** *или отправили в СИЗО, то ответственных за это должностных лиц могут отправить в места лишения свободы уже на срок до четырёх лет…»* (пост пользователя с ID 8436). Другим примером является включением прямой цитаты из Конституции РФ: *«…Если вас задержали за совершение какого-либо правонарушения, если вас просто пригласили как свидетеля, ничего не говорите. Всегда ссылайтесь на 51-ю статью Конституции: "Никто не обязан свидетельствовать против себя самого, своего супруга и близких родственников". Когда человека задерживают, особенно если он невиновен, его первая реакция – попытаться доказать свою невиновность…»* (пост пользователя с ID 8436).

Другие коэффициенты также указывают на умеренную или сильную связь при $p < 0.05$. Среди положительных коэффициентов выделим следующие: $r = 0.8461$ для группы «имя существительное-имя прилагательное» и $r = 0.951$ для группы «глагол-наречие». Умеренной связью с коэффициентом $r = 0.6573$ обладает группа «длина предложения-степень дистантизации». К группам с отрицательными коэффициентами относятся корреляции имен существительных и глаголов, а также корреляции имен прилагательных и наречий: $r = -0.7273$ и $r = -0.7062$ соответственно (рис. 32).

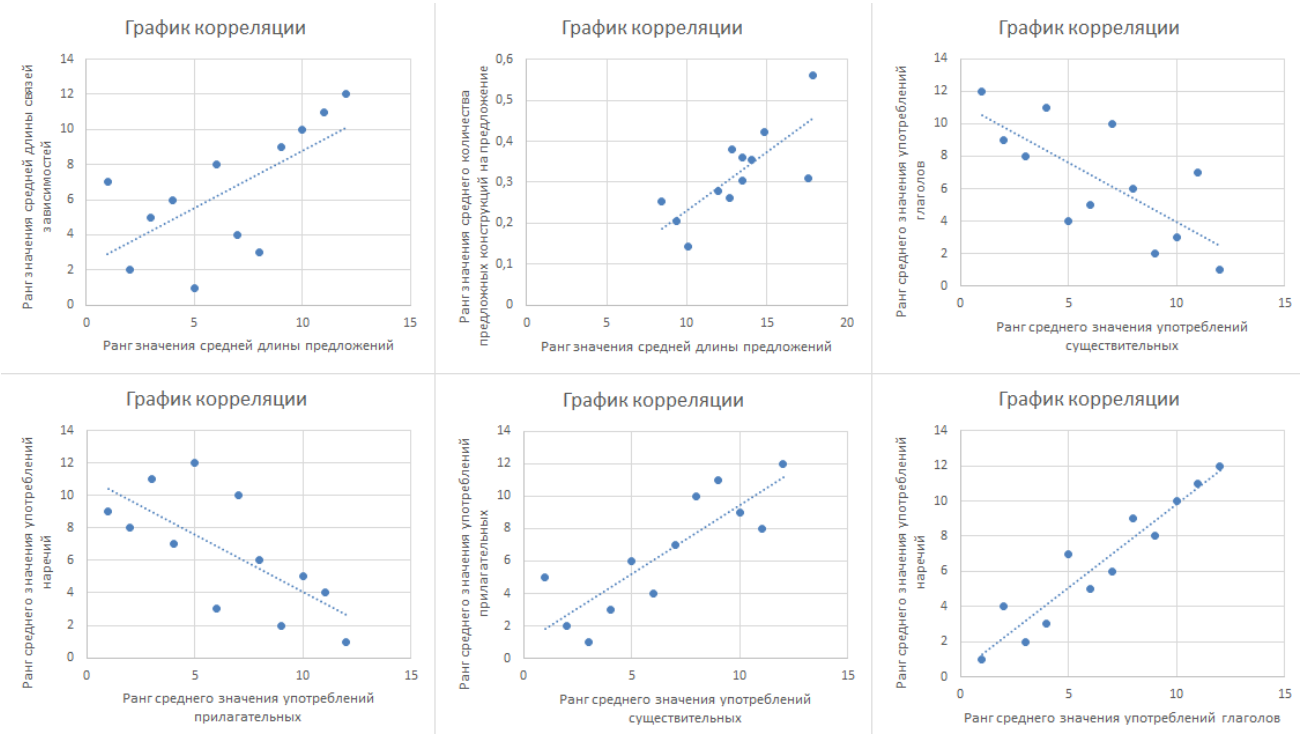

Рисунок 32 — Сводные данные о внутритекстовых корреляциях

3.7.13 Скрытое сообщество «Природа»

При анализе постов 39 пользователей коэффициенты корреляции в совокупном представлении и t-статистика значимы для четырех групп: «имя существительное-имя прилагательное», «глагол-наречие», «длина предложения-степень дистантизации» и «длина предложения-количество предложных конструкций». Все полученные данные демонстрируют прямую связь в диапазоне $r \in [0.3301; 0.6632]$: нижная граница – морфологический коэффициент глаголов и наречий, а верхняя граница – морфологический коэффициент имен существительных и имен прилагательных. Тематический компонент природы в постах используется для описания локаций, а активное использование модификаторов имен существительных и глаголов указывает на желание пользователя как можно ярче передать цветовую палитру реальной жизни: *«…У нас* ***крупная плоская*** *галька,* ***широкий*** *и* ***длинный*** *пляж без* ***железной*** *дороги (как известно именно* ***железная*** *дорога портит пляжи от Туапсе и до Сочи), никаких волнорезов - только* ***бескрайняя синяя*** *вечность со всех сторон (кроме той, с которой галька))…»* (пользователь с ID 49156), *«…****Сказочные*** *деревья, небо цвета ультрамарин и самое* ***Белое*** *в мире море - вот чудо похода в* ***зимнюю*** *сказку*

Колвицких тундр…» (пользователь с ID 40585). Возможность выражение экспрессии авторами также заключается в привлечении усложненных синтаксических конструкций. Так, в предложной конструкции в предложении *«…Растрескавшаяся земля впитает воду, солнце высушит листья и разогреет воздух* ***до привычных уже двадцати девяти градусов****…»* (пользователь с ID 49156) степень дистантизации между предлогом и существительным в форме родительного падежа равна четырем при общей длине предложения в 18 узлов синтаксического дерева.

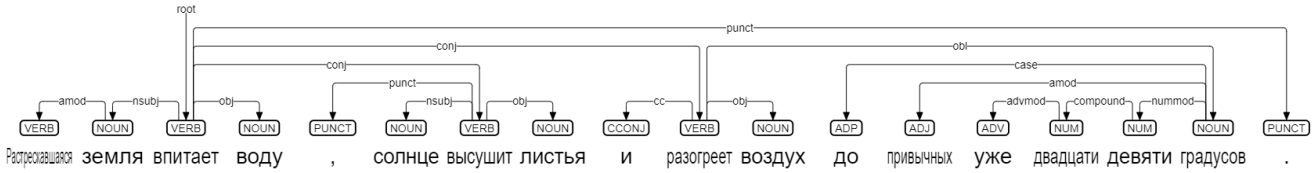

Рисунок 33 — Дерево зависимостей для предложения из поста пользователя с ID 49156

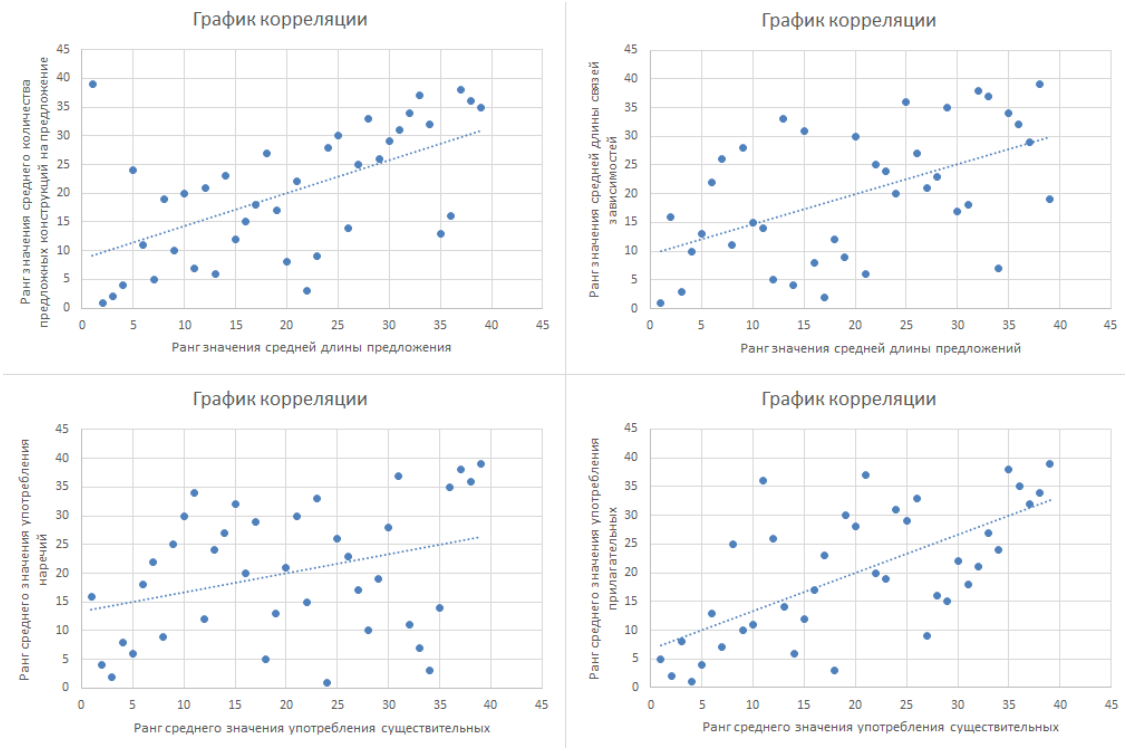

Рисунок 34 — Сводные данные о внутритекстовых корреляциях

3.7.14 Скрытое сообщество «Происшествие»

В результате вычислений было установлено, что в данном тематическое сообщество входят 37 пользователей. Для их постов было выделено четыре значимые корреляции из семи. Согласно количественной информации на рисунке

35, можно установить, что морфологические взаимосвязи имеют слабо-умеренную степень зависимости: коэффициент корреляции $r = 0.4637$, уровень значимости $p = 0.0038$ для существительных и прилагательных, а также $r = 0.3255$, уровень значимости $p = 0{,}049$ для глаголов и наречий. Синтаксические связи обладают бóльшими коэффициентами: $r = 0.6093$, $p = 0{,}0000627$ для группы «длина предложения-степень дистантизации» и $r = 0.8192$, $p = 0{,}000000000569$ для группы «длина предложения-количество предложных конструкций».

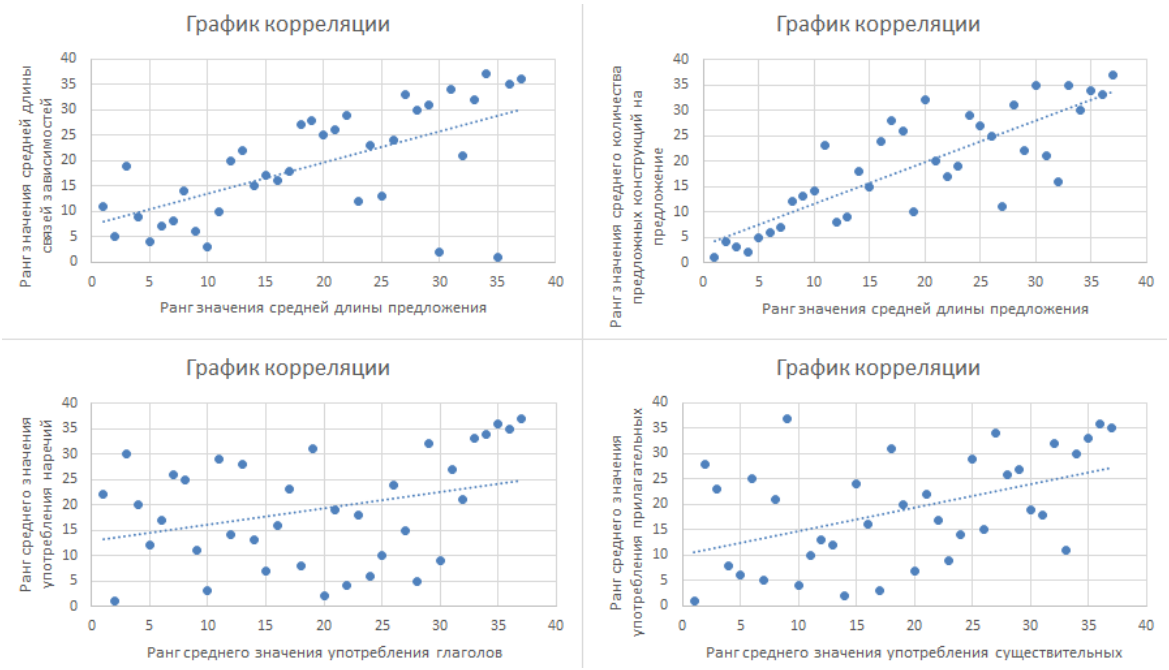

Рисунок 35 — Сводные данные о внутритекстовых корреляциях

3.7.15 Скрытое сообщество «Психология»

По данным рисунка 36 видно, что существует умеренная связь между длиной предложения в постах пользователей и средними значениями степени дистантизации и количеством предложных конструкций, $r = 0.6476$, $p = 0.0000000000215$ и $r = 0.6942$, $p = 0.000000000000173$ соответственно). Морфологические корреляции для групп «имя существительное-имя прилагательное» и «глагол-наречие» получили меньшие коэффициенты: $r = 0.4569$, $p = 0.000011$ и $r = 0.4512$, $p = 0.0000147$ соответственно. В целом критериальную валидность теста можно признать удовлетворительной. Существенные изменения наблюдается в группах «имя существительное-глагол» и «имя прилагательное-наречие»: как и в предыдущих разделах, для них характерны

обратные силы связи с низкими коэффициентами: $r = -0.3876$, $p = 0.000011$ и $r = -0.2244$, $p = 0.00025$ соответственно.

Рисунок 36 — Сводные данные о внутритекстовых корреляциях

3.7.16 Скрытое сообщество «Путешествие»

В данном сообществе наблюдается умеренная связь для двух морфологических и двух синтаксических групп при общем уровне значимости $p < 0.05$: для группы «имя существительное-имя прилагательное» значение $r = 0.6463$, для группы «глагол-наречие» – $r = 0.6427$, для группы «длина предложения-степень дистантизации» – $r = 0.4886$, для группы «длина предложения-количество предложных конструкций» – $r = 0.5253$. Еще двум парам значений были присвоены отрицательные коэффициенты со слабой связью: $r = -0.3515$ – для имен существительных и глаголов и $r = -0.3055$ – для имен прилагательных и наречий.

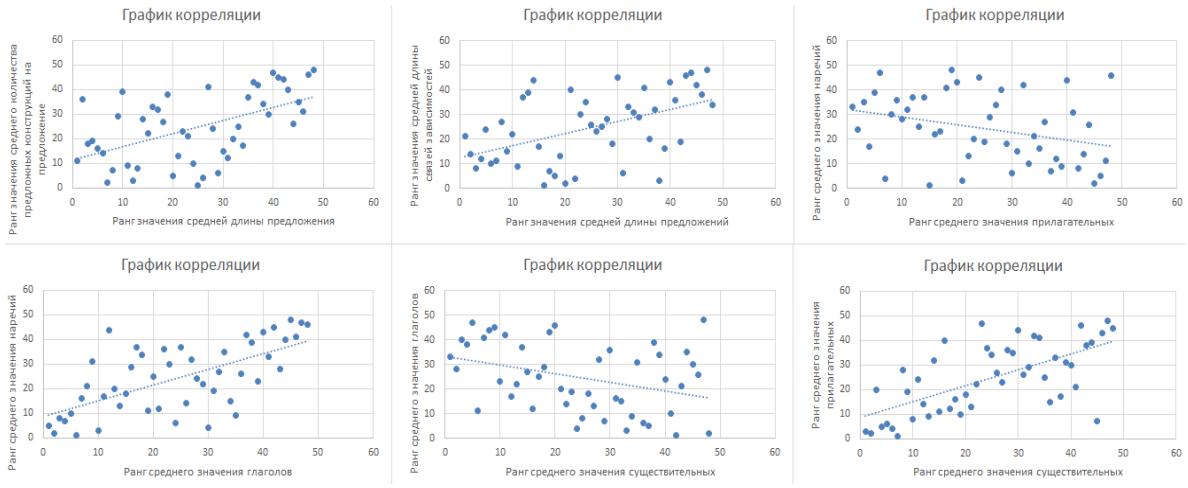

Рисунок 37 — Сводные данные о внутритекстовых корреляциях

3.7.17 Скрытое сообщество «Рабочий процесс»

В сообществе, в которое входят 36 пользователей, выявлено две значимые синтаксические корреляции и три значимые морфологические корреляции (рис. 38). Среди них лидирующую позицию занимает связь длины предложения с количеством предложных конструкций: $r = 0.6414$.

1. Пользователь с ID 172823: *«…Люблю повторять, что **за эти 10 лет** работал не **в одной компании**…»* (две предложные конструкции на 12 словоформ).
2. Пользователь с ID 726426: *«…Дачник приходит **на работу с синими кругами под глазами**…»* (три предложные конструкции на девять словоформ).
3. Пользователь с ID 77745: *«…Любая работа **в сфере управления репутацией в конечном итоге** направлена **на увеличение прибыли**…»* (три предложные конструкции на 13 словоформ).

Положительная сила корреляции также выявлена у глаголов и наречий-модификаторов: $r = 0.6086$. Значение $r = 0.6018$ установлено для длины предложения и степени дистантизации. Для имен существительных и имен прилагательных значение коэффициента корреляции находится приблизительно на том же уровне, что и для глаголов и наречий – $r = 0.5776$. Наконец, к отрицательным значениям относится $r = 0.3609$ для имен существительных и глаголов, оно находится в тех же диапазонах, что и в случае большинства проанализированных сообществ.

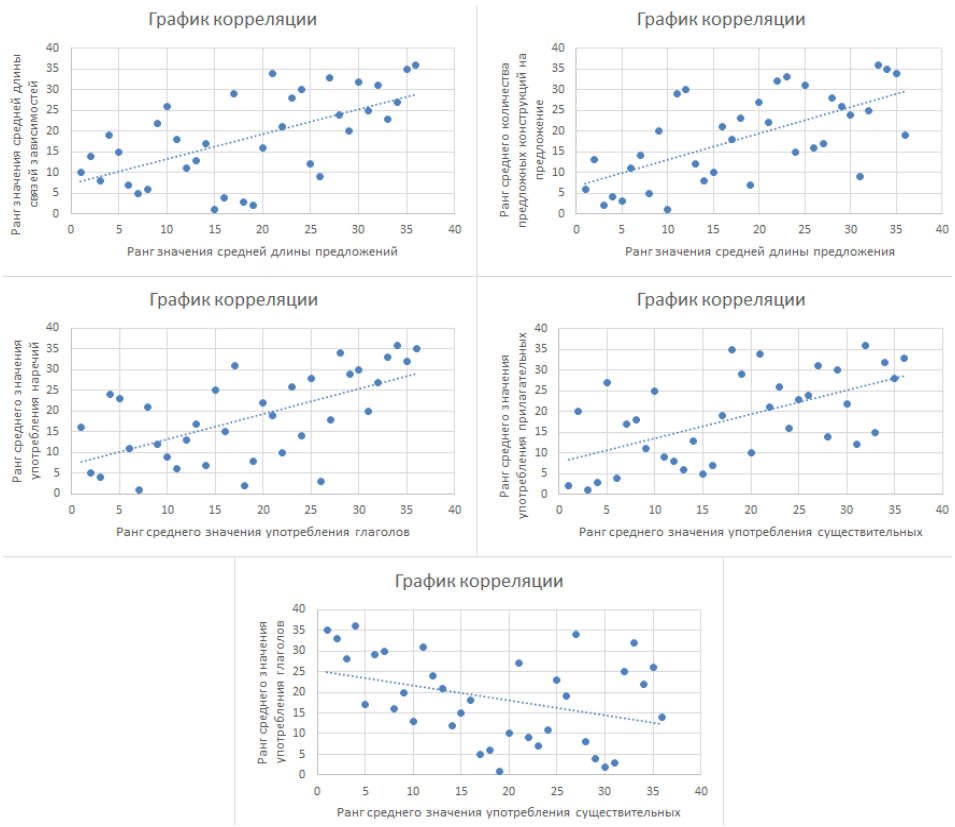

Рисунок 38 — Сводные данные о внутритекстовых корреляциях

3.7.18 Скрытое сообщество «Религия»

В сообществе из 14 участников отклоняющейся тенденцией от других значений из этой же категории является корреляция группы «коэффициент лексической плотности-коэффициент лексического разнообразия» – $r = -0.6201$, $p = 0.018$ (рис. 39). Это явление может объясняться тем, что при снижении количества лемм имен существительных, имен прилагательных, глаголов и наречий увеличивается количество лемм всех частей речи, в том числе и служебных. В постах религиозной тематики данное явление связано с тем, что пользователи часто используют характерные для священных писаний речевые обороты со служебными словами, например: *«…**И да простит** ему Господь все прегрешения **вольные же и невольные** и дарует ему Царствие Небесное…»* (предложение из поста пользователя с ID 184949). На данное значение корреляции также может повлиять преобладание разных видов местоимений в определенных отрывках постов: *«…Но, прежде чем **Он** успел донести людям **Своё** Послание полностью, **они** распяли **Его**…»* (предложение из поста пользователя с ID 30682).

Значения двух синтаксических корреляций ближе к единице, что указывает на сильную связь: $r = 0.9077$, $p = 0.0000073$ для группы корреляция «длина предложения-степень дистантизации» и $r = 0.8242$, $p = 0.00029$ для группы корреляция «длина предложения-количество предложных конструкций». Среди значимых морфологических корреляций выделена только одна – $r = 0.5824$, $p = 0.0289$ для группы «глагол-наречие».

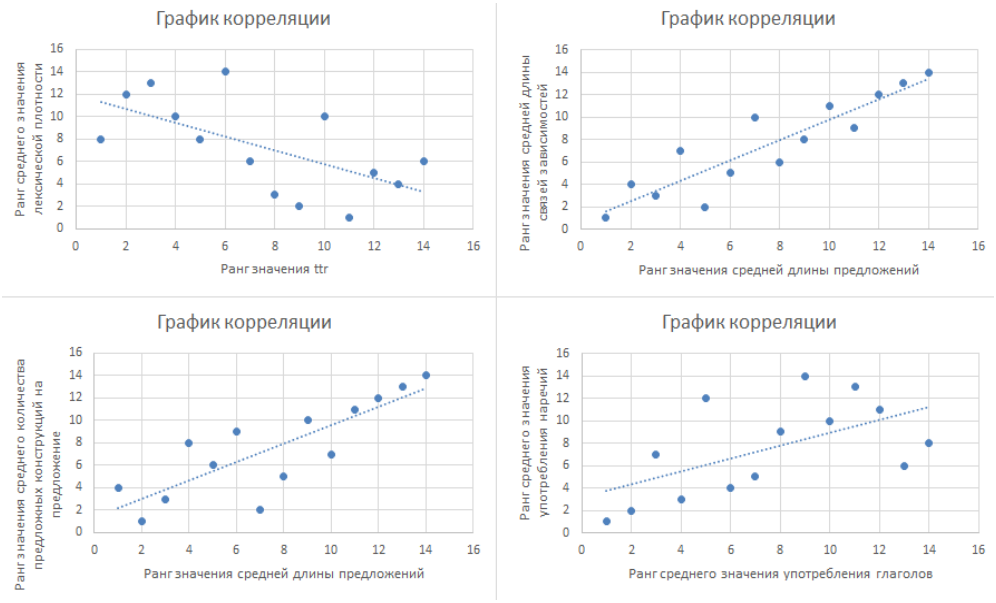

Рисунок 39 — Сводные данные о внутритекстовых корреляциях

3.7.19 Скрытое сообщество «Спорт»

В данном сообществе наблюдается умеренная связь для пяти морфосинтаксических и одной лексической групп при общем уровне значимости $p < 0.05$: для группы «имя существительное-имя прилагательное» значение $r = 0.4523$, для группы «глагол-наречие» – $r = 0.6423$, для группы «длина предложения-степень дистантизации» – $r = 0.5205$, для группы «длина предложения-количество предложных конструкций» – $r = 0.5611$, для группы «коэффициент лексической плотности-коэффициент лексического разнообразия» – $r = 0.2815$. Одной паре значений были присвоены отрицательные коэффициенты со слабой связью: $r = -0.2811$ – для имен существительных и глаголов.

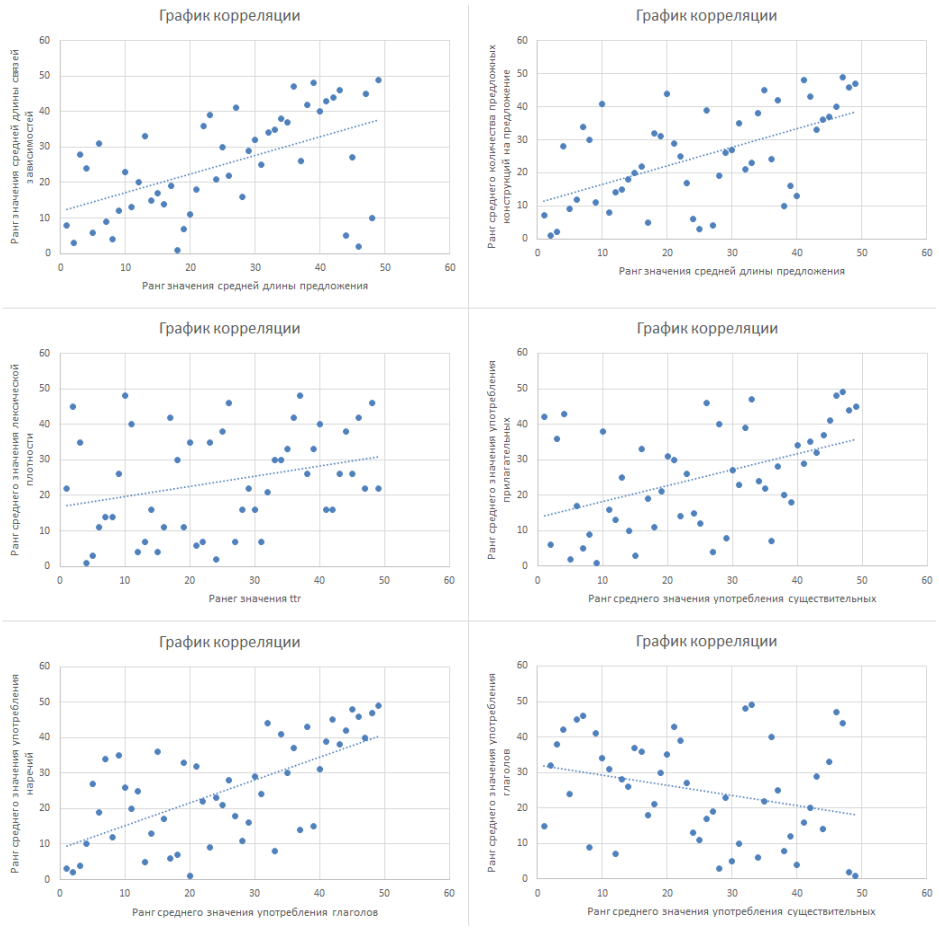

Рисунок 40 — Сводные данные о внутритекстовых корреляциях

3.7.20 Скрытое сообщество «Строительство и архитектура»

В данном сообществе было выделено всего две группы со значимыми корреляциями. Одна из них – прямая сила связи $r = 0.7$ при уровне значимости $p = 0.0165$ для имен существительных и прилагательных (рис. 41). Одной из причин для увеличения количества имен прилагательных при именах существительных может послужить использование в постах именованных сущностей, в состав которых входят атрибутивы: *«…Администрация **Пушкинского района** планирует завершить проектирование уже в апреле, а в июне приступить к первому этапу благоустройства…»* (предложение из поста пользователя с ID 636), *«…Сам дом был построен в 1896 году отцом Валерия Павловича и в 1956 году перенесён от береговой линии, из зоны затопления, при строительстве **Горьковской ГЭС**…»* (предложение из поста пользователя с ID 31287), *«…Это гостиница считалась наиболее престижной в **Армянском квартале**.…»* (предложение из поста пользователя с ID 2306).

Прямая сила связи $r = 0.8636$ при уровне значимости $p = 0.0006$ наблюдается также для глаголов и наречий (рис. 42). Также было рассчитано периферийное значение, которое не включено в итоговые лингвистические профили из-за уровня значимости, – $r = 0.5727$, $p = 0.0655$.

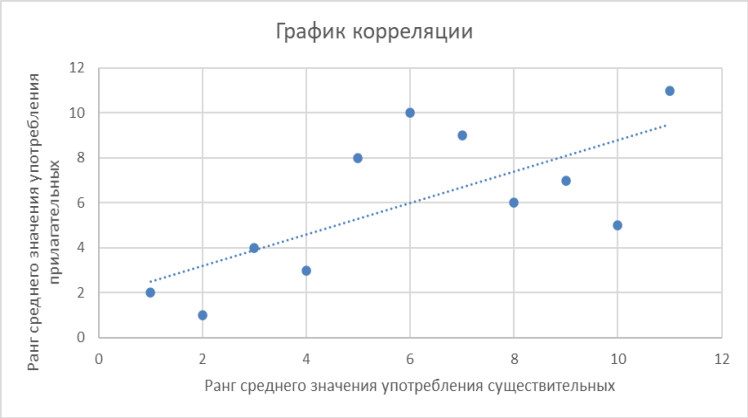

Рисунок 41 — Внутритекстовая корреляция средних значений употреблений имен существительных и имен прилагательных

Рисунок 42 — Внутритекстовая корреляция средних значений употреблений глаголов и наречий

3.7.21 Скрытое сообщество «Транспорт»

В данном сообществе, в которое входят 18 участников, наблюдается средняя сила прямой связи для четырех морфосинтаксических групп при общем уровне значимости $p < 0.05$: для группы «имя существительное-имя прилагательное» значение $r = 0.7193$, для группы «глагол-наречие» – $r = 0.6636$, для группы «длина предложения-степень дистантизации» – $r = 0.548$, для группы «длина предложения-количество предложных конструкций» – $r = 0.6058$. Для группы имен существительных и глаголов наблюдается обратная сила связи $r = -0.4427$,

однако периферийный уровень значимости $p = 0.0658$ не позволил включить данное значение в итоговый профиль сообщества.

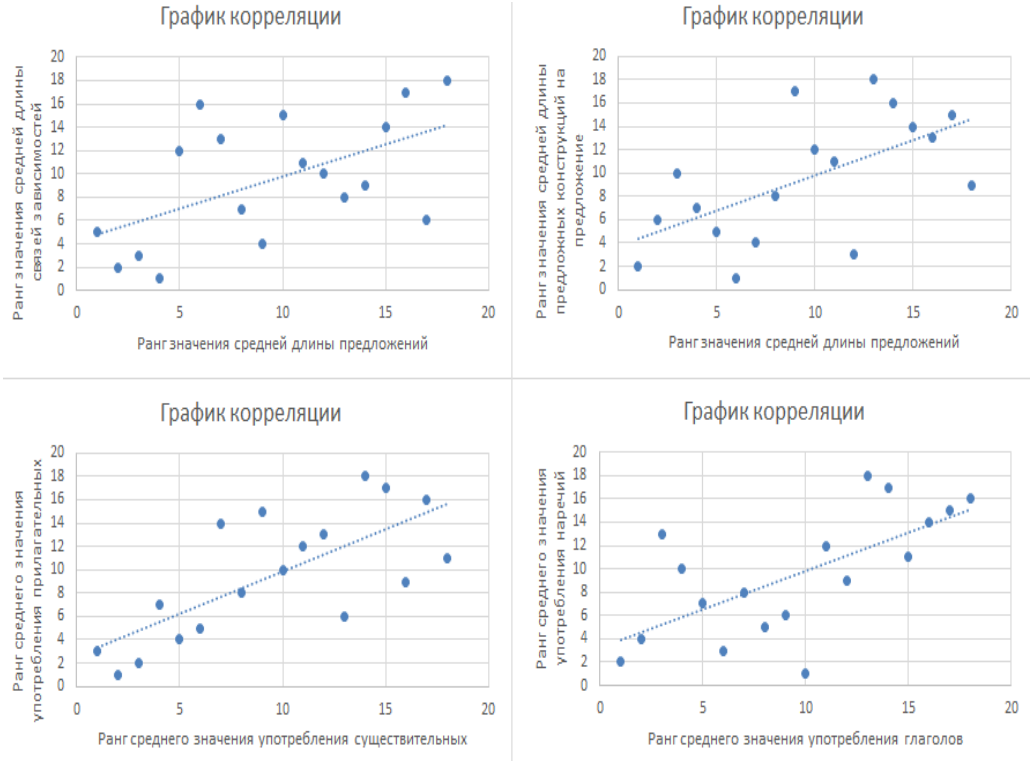

Рисунок 43 — Сводные данные о внутритекстовых корреляциях

3.7.22 Скрытое сообщество «Частная жизнь»

Второе по размеру скрытое сообщество в исследовательском корпусе охватывает посты, связанные с общедоступным описанием жизнедеятельности каждого отдельного лица. Сводные данные по корреляциям представлены на рисунке 44.

Рисунок 44 — Сводные данные о внутритекстовых корреляциях

Для этих текстов характерно использование эмоционально окрашенной лексики, часть из которой представлена именами прилагательными, что повлияло на появление прямой связи с именами существительными (третий ряд) при $r = 0.4537$ и $p = 0.0000000112$, например: *«…Я стала* ***медленней****. Гораздо. Ещё 3 года назад меня было не остановить – вечно* ***ужаленный*** *в жопу персонаж с* ***нескончаемым*** *потоком энергии…»* (предложение из поста пользователя с ID 2617). Прямая сила характерна и для группы «длина предложения-степень дистантизации» (первый ряд, правая колонка): $r = 0.6376$, $p = 8.40975E - 18$. На рисунке 45 в придаточной части предложения расстояние между подлежащим и сказуемым равна четырем узлам. Обратная связь наблюдается для имен существительных и глаголов (второй ряд, правая колонка): $r = -0.3672$, $p = 0.00000599$.

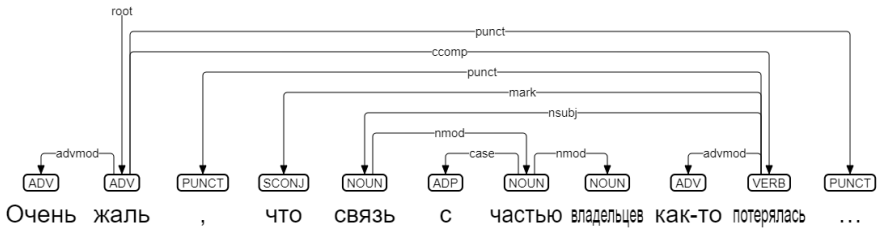

Рисунок 45 — Дерево зависимостей для предложения из поста пользователя с ID 10416

3.7.23 Скрытое сообщество «Эзотерика»

В последнем сообществе из 17 участников в результате вычислений оказалось, что всего одна корреляция из семи является значимой – это группа «длина предложения-степень дистантизации» с прямой силой связи $r = 0.4877$ при уровне значимости $p = 0.047$ (рис. 46).

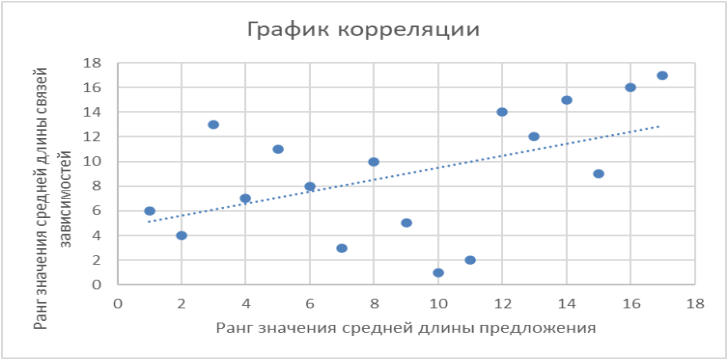

Рисунок 46 — Внутритекстовая корреляция средних значений длины предложений и длины структур зависимостей

Для пользователя с ID 2644 встречаются примеры инверсионной структуры предложения вида OVS с расширенным дополнением, что приводит к увеличению расстояния между главным и зависимым узлом, например, на рисунке 47 степень дистантизации между предикатом «является» и левостоящим зависимым дополнением «дверью» равняется трем при 12 узлах синтаксического дерева.

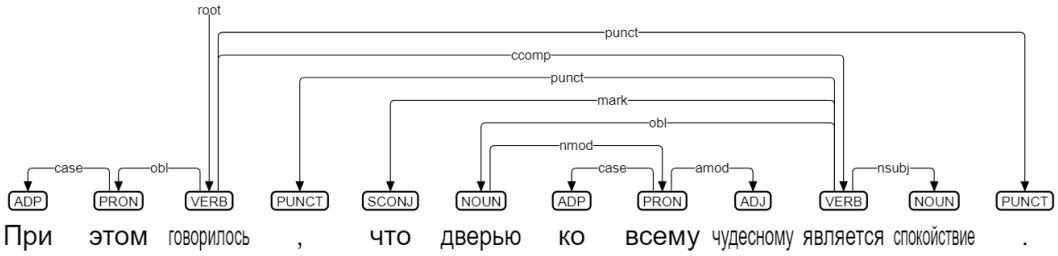

Рисунок 47 — Дерево зависимостей для предложения из поста пользователя с ID 2644

В постах с прямым порядком слов степень дистантизации может увеличиваться при подчинении нескольких зависимых дополнений при единой глагольной вершине. Так, в предложении из поста про медитацию пользователя с ID 9074 в главной части предложения расстояние между инфинитивом и косвенным дополнением увеличено за счет стоящего между ними прямого дополнения.

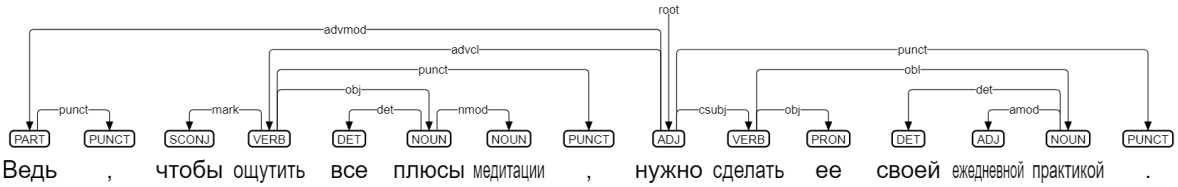

Рисунок 48 — Дерево зависимостей для предложения из поста пользователя с ID 9074

Наконец, увеличение степени дистантизации связано с правосторонним расширением подлежащих. Например, на рисунке 49 степень дистантизации между подлежащим и сказуемым в главной части предложения равняется четырем при общей длине предложения в 17 узлов синтаксического дерева.

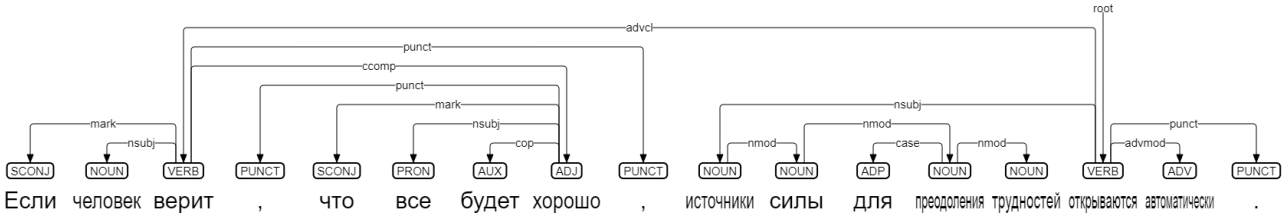

Рисунок 49 — Дерево зависимостей для предложения из поста пользователя с ID 51196

3.8 Кластерные группы лингвистических профилей

Полученные результаты свидетельствуют о вариациях стиля пользователей тематических скрытых сообществ на морфосинтаксическом и лексическом уровнях, а также о специфической организации постов социальных сетей. Итоговые количественные профили скрытых сообществ представлены в таблице 6, для которой введем условные обозначения: NV обозначает корреляцию «имя существительное-глагол»; NAdj обозначает корреляцию «имя существительное-имя прилагательное»; VAdv обозначает корреляцию «глагол-наречие»; AdjAdv обозначает корреляцию «имя прилагательное-наречие»; SenLin обозначает корреляцию «длина предложения-степень дистантизации»; LinPrep обозначает

корреляцию «длина предложения-количество предложных конструкций»; TtrDen обозначает корреляцию «коэффициент лексической плотности-коэффициент лексического разнообразия» [Мамаев, 2024а].

Таблица 6 — Лингвистические профили скрытых сообществ

Скрытое сообщество	**Морфологические корреляции**				**Синтаксические корреляции**		**Лексические корреляции**
	NV	**NAdj**	**VAdv**	**AdjAdv**	**SenLin**	**LinPrep**	**TtrDen**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Армия и государственная безопасность	—	0.7206	0.5245	-0.5196	0.7623	0.87	—
Астрономия	—	—	—	—	—	—	—
Бизнес, коммерция, экономика, финансы	-0.4182	0.4666	0.6419	—	0.6617	0.7257	—
Биология	—	—	—	—	—	—	—
География	—	—	—	—	—	—	—
Дом и домашнее хозяйство	-0.4784	0.4661	0.4008	—	0.6414	0.7081	—
Досуг, зрелища и развлечения	—	0.4044	0.6008	0.1742	0.4394	0.6342	—
Журналистика	—	—	—	—	—	—	—
Здоровье и медицина	-0.4228	0.5434	0.5979	—	0.5239	0.5481	0.286
Информатика	—	—	—	—	—	—	—
Искусство и культура	-0.1924	0.5135	0.6791	—	0.5946	0.7565	—
История	-0.416	—	0.4041	—	0.5197	0.7717	—
Легкая и пищевая промышленность	-0.7069	0.8473	0.6223	-0.5933	—	0.7069	—
Машиностроение	—	—	—	—	—	—	—
Наука и технологии	—	0.6084	—	—	0.6503	0.8392	—
Образование	-0.3633	0.2757	0.4711	—	0.4125	0.5233	0.2262
Политика и общественная жизнь	—	0.4856	0.6715	—	0.4304	0.7179	—
Право	-0.7273	0.8461	0.951	-0.7062	0.6573	0.8182	—
Природа	—	0.6632	0.3301	—	0.5225	0.5713	—
Производство	—	—	—	—	—	—	—
Происшествие	—	0.4637	0.3255	—	0.6093	0.8192	—
Психология	-0.3876	0.4569	0.4512	-0.2244	0.6474	0.6942	—
Путешествие	-0.3515	0.6463	0.6427	-0.3055	0.4886	0.5253	—
Рабочий процесс	-0.3609	0.5776	0.6086	—	0.6018	0.6414	—
Религия	—	—	0.5824	—	0.9077	0.8241	-0.6201

Продолжение таблицы 6

1	2	3	4	5	6	7	8
Социология	—	—	—	—	—	—	—
Спорт	-0.2811	0.4523	0.6423	—	0.5205	0.5611	0.2815
Строительство и архитектура	—	0.7	0.8636	—	—	—	—
Техника	—	—	—	—	—	—	—
Транспорт	—	0.7193	0.6636	—	0.548	0.6058	—
Филология	—	—	—	—	—	—	—
Философия	—	—	—	—	—	—	—
Частная жизнь	-0.3672	0.4537	—	0.5034	0.6376	0.7311	—
Эзотерика	—	—	—	—	0.4877	—	—

Вышеуказанные данные можно преобразовать в данные о близости сообществ по лингвистических параметрам, применив кластерный анализ, что позволит выделить группы, в которых пользователи используют близкие языковые средства. С помощью компьютерной программы *Orange*[53] составлен следующий пайплайн – непрерывный автоматизированный процесс обработки количественных данных. В среду была загружена xlsx-таблица с данными о 23 проанализированных сообществах. Оставшиеся 11 сообществ без корреляций в кластерный анализ не включались. В лингвистических профилях возникла проблема пропущенных значений, для ее решения использовался метод восстановления данных *model-based imputation*, так как он доказал свою эффективность в ряде лингвистических исследований [Kekez, 2021; Chakraborty, Kim, Sudhir, 2022]. Этот метод строит модель для прогнозирования отсутствующего значения на основе значений других атрибутов, для каждого атрибута создается отдельная модель. По умолчанию используется алгоритм 1-NN, который берет значение из наиболее похожего примера данных. Для создания дендрограммы необходимо рассчитать расстояния между рядами данных, в связи с чем было применено евклидово расстояние, была получена следующая матрица расстояний (рис. 50). Значения, представленные красными оттенками, указывают на большую степень дистантизации сообществ друг от друга на основании лингвистических параметров, таким образом, они потенциально будут принадлежать к разным кластерам. Значения синих оттенков

[53] https://orangedatamining.com/

указывает на потенциальную возможность принадлежности сообществ к одному кластеру на основании близости лингвистических признаков.

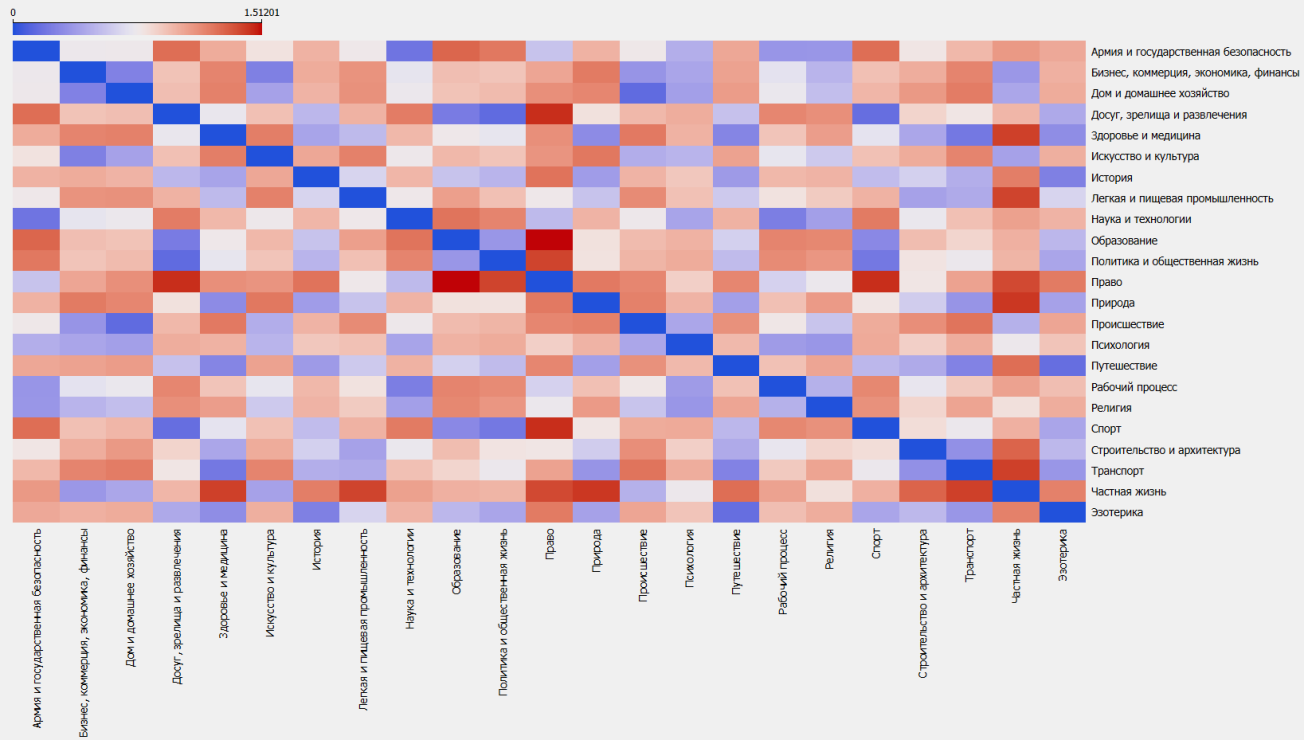

Рисунок 50 — Матрица расстояний для лингвистических профилей

Наконец, 23 профиля были разбиты на три кластера методом Варда (рис. 51), каждый элемент которого был оценен с точки зрения Silhouette Score (рис. 52): значение оценки показывает, близок ли лингвистический профиль к своей группе. Отсутствие отрицательных значений указывает на то, что шанс потенциального отнесения лингвистических профилей к другим кластерам низок.

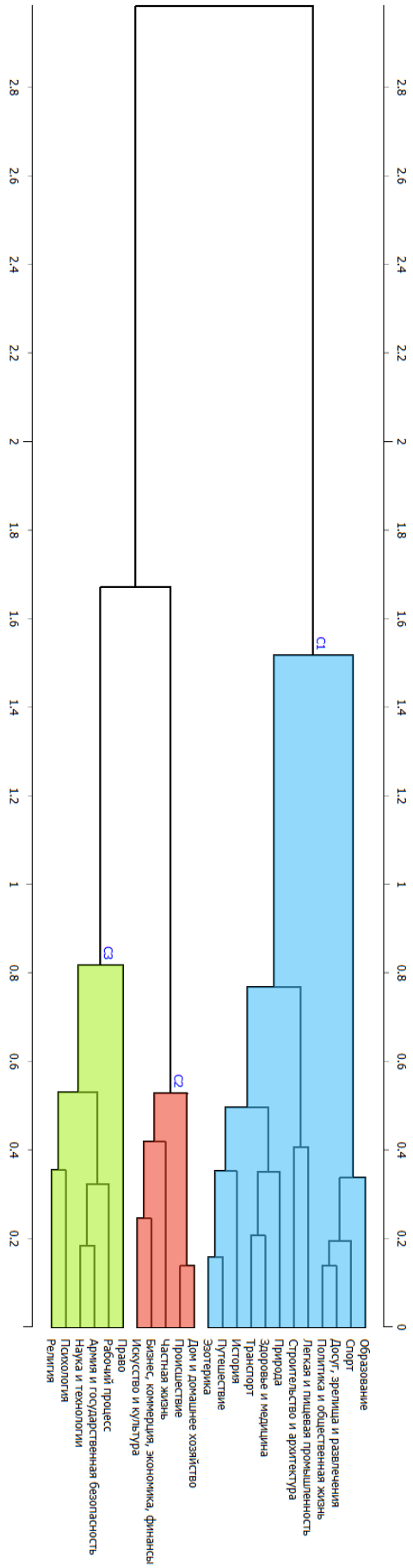

Рисунок 51 — Кластеры лингвистических профилей

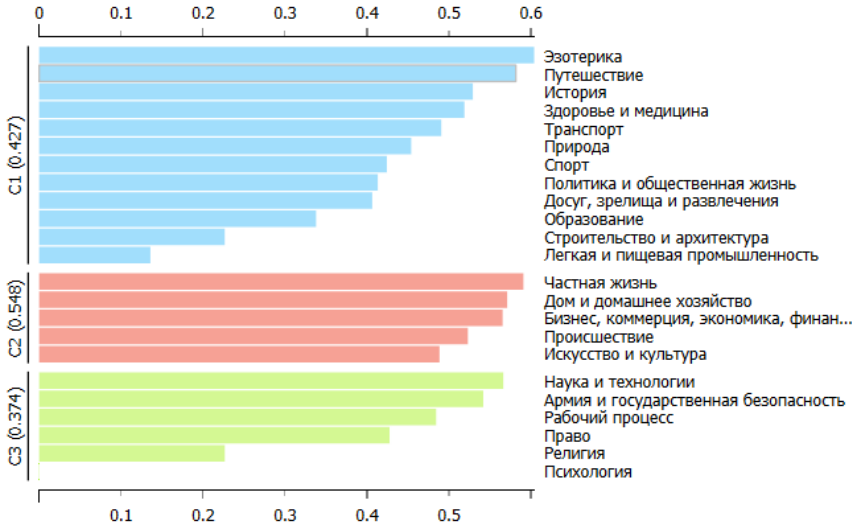

Рисунок 52 — Оценка степени принадлежности профилей к кластерам

Приведенные на рисунке 53 данные свидетельствуют о том, что разница между морфологическими корреляциями «имена существительные-глаголы» и «глаголы-наречия» не существенна, а в случае медиан (желтая линия) для корреляции «глаголы-наречия» разница минимальна. Очевидные различия между кластерами начинают проявляться на синтаксическом уровне, что видно благодаря расположению асимметрии «ящиков» (рис. 54).

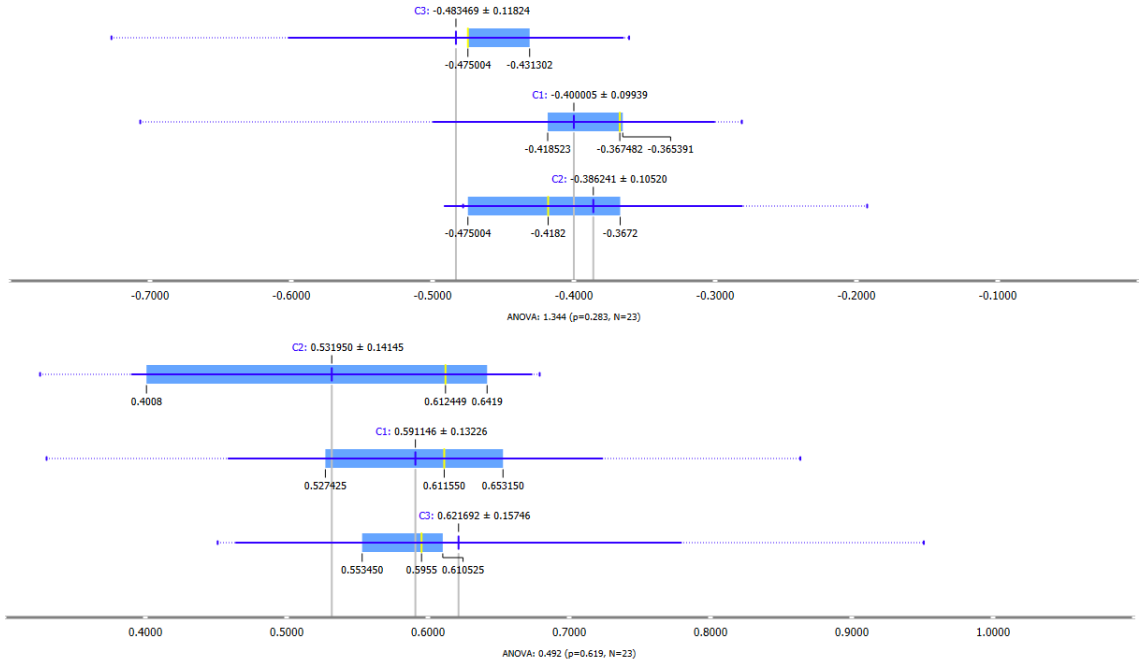

Рисунок 53 — Диаграмма размаха для корреляций «имена существительные-глаголы» и «глаголы-наречия»

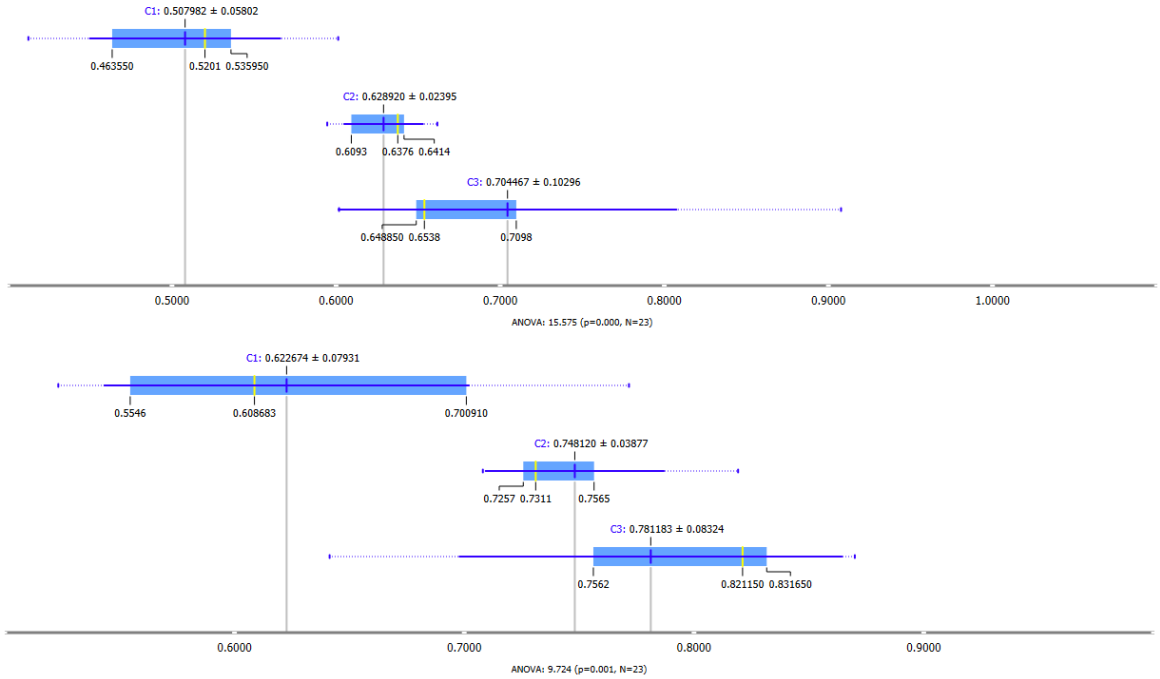

Рисунок 54 — Диаграмма размаха для синтаксических корреляций

Выводы по третьей главе

На основании модели скрытых сообществ, полученной с помощью семантических анализаторов, и результатов применения разработанной процедуры лингвистического профилирования можно сделать ряд выводов.

1. Использование предобученных языковых моделей и ручное внедрение параметра «автор» позволяет настроить алгоритмы тематического моделирования на выделение узконаправленных тематических списков.

2. «Густой» центр графовой модели скрытых сообществ отображает способность пользователей создавать политематические тексты, что отражает их возможную заинтересованность различными областями жизни, в то время как «разреженная» периферия графа указывает на то, что пользователи, скорее всего, нацелены на порождение монотематических текстов.

3. Математические методы оперируют формальными показателями при выделении сообществ и не учитывают лингвистические характеристики текстов пользователей, что может привести к искажению реального числа неявных групп.

4. Несмотря на то, что пользователи рассматриваются как некоторый цифровой образ, некоторые извлеченные социально-демографические характеристики пользователей соотносятся с реальными данными: например,

большой процент пользователей, посты которых включены в модель, проживает в Санкт-Петербурге и Москве, что может указывать на реальную густонаселенность регионов (см. данные Росстата[54]).

5. Введение проверки значимости выявленных корреляций показало, что лексические параметры практически не представлены в лингвистических профилях пользователей, что приводит к затруднению количественной оценки коммуникативно-письменных навыков.

6. Кластерный анализ показал, что создаваемые посты пользователями скрытых сообществ с точки зрения синтаксиса разнообразнее, чем с точки зрения морфологии.

54 https://rosstat.gov.ru/folder/313/document/166784#:~:text=%D0%A1%D0%B0%D0%BC%D1%8B%D0%BC%D0%B8%20%D0%B3%D1%83%D1%81%D1%82%D0%BE%D0%BD%D0%B0%D1%81%D0%B5%D0%BB%D0%B5%D0%BD%D0%BD%D1%8B%D0%BC%D0%B8%20%D1%80%D0%B5%D0%B3%D0%B8%D0%BE%D0%BD%D0%B0%D0%BC%D0%B8%20%D0%A0%D0%BE%D1%81%D1%81%D0%B8%D0%B8%20%D0%BE%D1%81%D1%82%D0%B0%D1%8E%D1%82%D1%81%D1%8F,%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5%20%D0%BD%D0%B0%20%D0%BE%D0%BA%D1%82%D1%8F%D0%B1%D1%80%D1%8C%202021%20%D0%B3%D0%BE%D0%B4%D0%B0

ЗАКЛЮЧЕНИЕ

В современном мире социальных сетей стремительно увеличивается активность пользователей при обсуждении как повседневных, так и профессиональных вопросов, что приводит к появлению более тесных, но не всегда явных взаимосвязей между пользователями. Обнаружение скрытых сообществ и свойственных им характеристик помогает глубже понять окружающее общество, раскрыть сущность сложных социальных явлений и попытаться охарактеризовать пользователей с точки зрения языковых навыков.

В данном диссертационном исследовании была достигнута цель – разработана процедура лингвистического профилирования сообществ в социальных медиа на основе пользовательских текстов, которые представлены в корпусе русскоязычных постов. Получены следующие результаты.

1. Комбинирование лингвистических и экстралингвистических элементов в текстах социальных сетей усложняет их автоматическую обработку, в связи с чем необходимо обращаться к инструментам компьютерной лингвистики, которые минимизируют потерю данных.

2. Для создания корпуса текстов, предназначенного для идентификации скрытых сообществ, нужно учесть следующие критерии: использование только письменных (клавиатурно-опосредованных) текстов, сбалансированность по параметру пола и времени публикации постов, парсинг данных из единой социальной сети, отсутствие общих друзей в социальных сетях у пользователей, чьи тексты формируют корпус.

3. Наиболее оптимальным способом обработки постов является использование нескольких модулей обработки текстов, некоторые из которых основаны на нейросетевых архитектурах и ранее апробированы на разножанровых корпусах.

4. При обработке исследовательского корпуса необходимо вручную постредактировать результаты, что связано с нестандартными способами оформления постов: внедрение символов латиницы в слова, написанные на кириллице, комбинация пунктуационных символов и слов и пр.

5. Использование методов контекстуализированного тематического моделирования позволяет восполнить пробелы в современной теории выявления скрытых сообществ.

6. На основании расчета коэффициента модуляции (формальный подход) установлено, что в итоговой модели выделено четыре сообщества. На основании экспертной разметки тематических моделей (лингвистический подход) выявлено 34 сообщества, из которых 23 подвергнуты дальнейшей процедуре лингвистического профилирования.

7. Для процедуры лингвистического профилирования отобраны параметры на трех языковых уровнях: морфологическом, синтаксическом и лексическом, с помощью инструментов лингвостатистического анализа текстов извлечена количественная информация об использовании параметров.

8. Во время процедуры лингвистического профилирования на основании полученных количественных данных рассчитаны внутритекстовые корреляты. Исследуемые пары выборок проверялись на нормальность распределения, на основании чего в дальнейшем выбирался необходимый коэффициент корреляции. Полученные лингвистические профили, представленные в форме кортежа, были подвергнуты многомерным методам анализа – кластеризации и дисперсионному анализу.

Таким образом, разработанный в диссертации корпус стал эмпирической базой для выявления и интерпретации 23 лингвистических профилей скрытых сообществ, при этом ни одно сообщество не было представлено полным набором из семи корреляций. Максимальное количество значимых корреляций в профиле сообщества достигало шести, а минимальное – одного, что может быть связано с числом данных в анализируемой выборке. Подобное лингвистическое профилирование в исследовании проведено для того, чтобы создать функциональную модель, которая в действительности отражает текущие языковые тенденции в текстах социальных сетей, объединенных единым тематическим компонентом. Подобная модель может найти практическое применение при создании систем автоматической модерации групп и отслеживания тенденций

среди пользователей, на основании чего рекламные группы смогут модифицировать лексические конструкции и синтаксис своих постов для привлечения большего количества клиентов. Выдвинутая в диссертации **гипотеза подтвердилась**.

Проведенное исследование имеет высокую практическую значимость для специалистов в области социолингвистики и медиаисследований, которые решают задачи, связанные с оптимизацией архитектуры социальных сетей и СМИ. Конечно, представляется важным продолжить исследование в следующих направлениях.

1. Проведение экспериментов по созданию лингвистических профилей скрытых сообществ на материале других русскоязычных социальных сетей: Одноклассники, LiveJournal и др.

2. В текущей работе была представлена статическая информация о тематике постов и связях между пользователями социальных сетей. Выявления изменений в постах пользователей на морфологическом, синтаксическом и лексическом уровнях можно добиться при внедрении алгоритмов динамического тематического моделирования.

3. Проведение работ по дальнейшей автоматизации алгоритма: замена процедуры ручной разметки тем на автоматическую за счет привлечения современных поисковых систем и векторных моделей разножанровых русскоязычных корпусов. В частности, некоторые из них представлены на специализированных платформах, например, RusVectōrēs[55].

[55] https://rusvectores.org/ru/

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Аванесян Н. Л. Выявление значимых признаков противоправных текстов / Н. Л. Аванесян, Ф. Н. Соловьев, Е. А. Тихомирова, А. М. Чеповский // Вопросы кибербезопасности. – 2020. – № 4 (38). – С. 76-84.

2. Алексеева А. А. Исследование текстов в социальной сети «Вконтакте» в лингводидактическом аспекте / А. А. Алексеева // Лингвокультурология. – 2014. – № 8. – С. 6-10.

3. Алисов Е. А. Технологии оценки качества освоения новых модулей педагогической магистратуры / Е. А. Алисов // Современные проблемы науки и образования. – 2015. – № 2-1. – С. 472-472.

4. Апажева Л. Т. Модель языковой личности субъекта виртуальной коммуникации / Л. Т. Апажева // Актуальные проблемы филологии и педагогической лингвистики. – 2014. – №. 16. – С. 138-144.

5. Баранский В. А. Учебно-методический комплекс дисциплины «Графы и матроиды» / В. А. Баранский, В. В. Расин. – Екатеринбург: Уральский государственный университет им. А. М. Горького, 2008. – 157 с.

6. Бодрова Т. Определение коэффициента ранговой корреляции частей речи в русских и чувашских газетных текстах / Т. Бодрова, Н. Тукмакова // Мовознавчий вісник. – 2012. – №. 14-15. – С. 374-382.

7. Бодулева А. Р. Лингвистические особенности СМС-сообщений английского языка / А. Р. Бодулева, А. З. Зарипова // Инновационная наука. – 2016. – № 2-5 (14). – С. 76-78.

8. Быкова А. П. Оценка эмоциональной окраски словосочетаний в постах социальной сети «Вконтакте» методами машинного обучения / А. П. Быкова. – СПб: Санкт-Петербургский государственный университет, 2023. – 83 с.

9. Волобуев И. В. Языковые средства выразительности рекламного текста на английском языке / И. В. Волобуев // Вестник Адыгейского государственного университета. Серия 2: Филология и искусствоведение. – 2013. – № 3 (126). – С. 37-41.

10. Воронин А. Н. Взаимосвязь сетевых характеристик и субъектности сетевых сообществ в социальной сети Твиттер / А. Н. Воронин, Ю. В. Ковалева, А. А. Чеповский // Вопросы кибербезопасности. – 2020. – № 3 (37). – С. 40-57.

11. ГОСТ 7.79-2000 Правила транслитерации кирилловского письма латинским алфавитом. – Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2000. – 22 с.

12. Градосельская Г. В. Картирование политически активных групп в Фейсбуке: динамика 2013-2018 гг. / Г. В. Градосельская, Т. Е. Щеглова, И. А. Карпов // Вопросы кибербезопасности. – 2019. – № 4 (32). – С. 94-104.

13. Гридина Т. А. Языковая игра в современной интернет-коммуникации: метаязыковой аспект / Т. А. Гридина, С. С. Талашманов // Политическая лингвистика. – 2019. – №. 3. – С. 31-37.

14. Гущин А. Н. Основы представления знаний: учеб. пособие / А. Н. Гущин. – СПб: Балт. гос. техн. ун-т, 2007. – 31 с.

15. Долгих Е. А. Анализ возможностей использования цифрового следа в системе высшего образования / Е. А. Долгих, Т. А. Першина // Тенденции развития науки и образования. – 2021. – № 76-2. – С. 10-16.

16. Задорожный В. Н. О неоднородной структуре социальных сетей / В. Н. Задорожный, Е. Б. Юдин // Омский научный вестник. – 2017. – №. 2 (152). – С. 91-96.

17. Захаров В. П. Корпусная лингвистика: учебник. 3-е изд., перераб. / В. П. Захаров, С. Ю. Богданова. – СПб.: Изд-во С.-Петерб. ун-та, 2020. – 234 с.

18. Иноземцева Н. В. Парцелляция как основная синтаксическая модель заголовков англоязычных статей по методической / Н. В. Иноземцева // Вестник Оренбургского государственного университета. – 2011. – №11 (130). – С. 114–118.

19. Кагиров И. А. Мультимедийная база данных жестов русского жестового языка в трехмерном формате / И. А. Кагиров, Д. А. Рюмин, А. А. Аксёнов, А. А. Карпов // Вопросы языкознания. – 2020. – №. 1. – С. 104-123.

20. Каинова М. М. Коммуникативно-письменный компетенции и письменная речь в теоретико-методическом и дидактическом аспекте в российской

науке и системе образования / М. М. Каинова // European Social Science Journal. – 2018. – № 7-1. – С. 322-331.

21. Кан Е. В. Хэштеги как новое лингвистическое явление / Е. В. Кан // Филологический аспект. – 2017. – № 1. – С. 91-98.

22. Кириченко Л. О. Обнаружение киберугроз с помощью анализа социальных сетей / Л. О. Кириченко, Т. А. Радивилова, А. Барановский // International Journal “Information Technologies & Knowledge”. – 2017. – № 11(1). – С. 23-48.

23. Конюшкевич М. И. Преобразование предложно-падежной синтаксемы в предикативную единицу: корреляция предлога и показателя связи сложного предложения / М. И. Конюшкевич // Лінгвістичні студії. – 2013. – № 26. – С. 93-99.

24. Крейн Д. Социальная структура группы ученых: проверка гипотезы о «невидимом колледже» / Д. Крейн // Коммуникация в современной науке. – М., 1976. – С. 183-218.

25. Крылова М. Н. Способы выражения эмоций в социальных сетях / М. Н. Крылова // Электронный научно-практический журнал «Филология и литературоведение». – 2016. – №1. – С. 78-84.

26. Крылова М. Н. Язык современного интернет-общения (на материале интеллектуального контента социальной сети" ВКонтакте") / М. Н. Крылова // Актуальные проблемы филологии и педагогической лингвистики. – 2019. – № 1. – С. 128-137.

27. Кувшинская Ю. М. Аббревиация в речи интернет-форумов / Ю. М. Кувшинская // Современный русский язык в интернете. – 2014. – С. 23-38.

28. Кутыркин А. В. Кластерный анализ: Методические указания / А. В. Кутыркин, А. В. Сёмин. – Переиздание. – М.: МИИТ, 2009. – 22 с.

29. Литвинова Т. А. Профилирование автора письменного текста / Т. А. Литвинова // Язык и культура. – 2013. – № 3 (23). – С. 64-72.

30. Лихачёва Ю. С. Цифровая личность: понятие и критерии описания / Ю. С. Лихачёва // Экономика и общество: перспективы развития. – 2020. – С. 245-248.

31. Малафеев О. А. Математическое моделирование задач экономической конкуренции по выявлению скрытых сообществ в социальной сети / О. А. Малафеева, С. А. Щеникова, О. И. Скворцова // Информационные технологии в образовании. – 2021. – С. 167-172.

32. Маллинз Н. Ч. Анализ содержания неформальной коммуникации между биологами / Н. Ч. Маллинз // Коммуникация в современной науке. – М., 1976. – С. 239-260.

33. Мамаев И. Д. Адаптация заимствованных слов к русской морфологии / И. Д. Мамаев, А. А. Зайцева // International Journal of Advanced Studies in Language and Communication. – 2019. – №2. – С. 106–113.

34. Мамаев И. Д. К вопросу о построении модели скрытых сообществ с помощью контекстуализированных тематических моделей / И. Д. Мамаев // Тезисы докладов 50-й Международной научной филологической конференции имени Людмилы Алексеевны Вербицкой. – 2022 (2022a). – С. 243.

35. Мамаев И. Д. Кластерный анализ лингвистических профилей скрытых сообществ / И. Д. Мамаев // Филологические науки. Вопросы теории и практики. – 2024 (2024a). – Т. 17. – Вып. 5. – С. 1739-1747.

36. Мамаев И. Д. Лингвистические особенности обработки текстов социальных сетей при построении модели скрытых сообществ / И. Д. Мамаев // Инновационные технологии и технические средства специального назначения: Труды четырнадцатой общероссийской научно-практической конференции. – Т. 2. – 2022 (2022b). – С. 312-315.

37. Мамаев И. Д. Лингвистические параметры для идентификации скрытых сетевых сообществ / И. Д. Мамаев, О. А. Митрофанова // Terra Linguistica. – 2024. – Т. 15. – №. 1. – С. 102-115.

38. Мамаев И. Д. Лингвистические профили скрытых сообществ: морфосинтаксический аспект / И. Д. Мамаев // Филологические науки. Вопросы теории и практики. – 2024 (2024b). – Т. 17. – Вып. 4. – С. 1155-1162.

39. Мартыненко Г. Я. Основы стилеметрии: учеб.-метод. пособие / Г. Я. Мартыненко, А. О. Гребенников. – СПб.: Изд-во С.-Петерб. ун-та, 2018. – 27 с.

40. Масликова О. С. Языковые особенности общения в Интернет-пространстве / О. С. Масликова // Инновационная наука. – 2019. – № 9. – С. 69-72.

41. Матусевич А. А. Общение в социальных сетях: прагматический, коммуникативный, лингвостилистический аспекты характеристики : дис. ... канд. филол. наук : 10.02.01 / Матусевич Александра Александровна. – Киров, 2016. – 190 с.

42. Мейлахс П. А. Онлайновое сообщество СПИД-диссидентов в социальной сети «ВКонтакте»: структура и риторические стратегии / П. А. Мейлахс, Ю. Г. Рыков // XV апрельская международная научная конференция по проблемам развития экономики и общества: в 4-х книгах. Отв. ред.: Е. Г. Ясин. – Кн. 3. – М.: Издательский дом НИУ ВШЭ. – 2015. – С. 137-146.

43. Минаев В. А. Как найти следы экстремизма в социальных медиа / В. А. Минаев // Противодействие терроризму и экстремизму в информационных сферах: сборник научных статей Всероссийской конференции. – 2022. – С. 15-19.

44. Митрофанова О. А. Динамическое тематическое моделирование русскоязычного корпуса юридических документов / О. А. Митрофанова, М. М. Атугодаге // Terra Linguistica. – 2023. – Т. 14. – № 1. – С. 70-87.

45. Митягин С. А. Исследование социальных сетей Интернет на предмет выявления сопутствующих интересов лиц, склонных к наркомании / С. А. Митягин, А. В. Якушев, А. В. Бухановский // Международный научно-исследовательский журнал. Технические науки. – 2012. – Вып. 6(6). – С. 59-64.

46. Некрасова Э. В. Анализ текста на соответствие заданной теме с применением методов машинного обучения / Э. В. Некрасова, П. Ю. Гусев // Научный аспект. – 2022. – №4. – Т. 1. – С. 100-110.

47. Никитин М. В. Полисемия на пределе (широкозначность) / М. В. Никитин // Язык. Человек. Общество. Междунар. сб. науч. тр. (к 60-летию профессора В. Т. Малыгина). – СПб.-Владимир: РГПУ им. А.И.Герцена, ВГПУ, 2005. – С.102-113.

48. Николенкова Н. В. Письменная коммуникация современных школьников как отражение уроков русского языка в средней школе /

Н. В. Николенкова // Лингвистика и школа — III. Материалы Всероссийской научно–практической конференции. – 2007. – С. 184-192.

49. Нокель М. А. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами / М. А. Нокель, Н. В. Лукашевич // Вычислительные методы и программирование. – 2015. – Т. 16. – С. 215-234.

50. Оболенский Д. М. Реализация классификатора групп в социальных сетях с помощью рекуррентных и свёрточных нейронных сетей / Д. М. Оболенский, В. И. Шевченко, О. В. Ченгарь, Е. Н. Мащенко, А. С. Соина // Экономика. Информатика. – 2021. – Т. 48. – №. 1. – С. 100-115.

51. Орехов Б. В. Метрическое и лексическое разнообразие в стихах А.А. Вознесенского / Б. В. Орехов // Труды института русского языка им. В.В. Виноградова. – 2022. – С. 50-58.

52. Печенкин В. В. Сетевые методы исследования виртуальных сообществ / В. В. Печенкин, В. В. Зайонц // Теория и практика общественного развития. – 2011. – №. 5. – С. 71-75.

53. Полидовец Н. И. О некоторых синтаксических особенностях современной интернет-коммуникации // Вестник Нижегородского университета им. Н.И. Лобачевского. – 2020. – №. 2. – С. 300-305.

54. Попов В. А. Выделение неявных пересекающихся сообществ на графе взаимодействия Telegram-каналов с помощью «метода Галактик» / В. А. Попов, А. А. Чеповский // Труды института системного анализа российской академии наук. – 2022. – Т. 72. – № 4. – С. 39-50.

55. Попова Д. А. Цифровая личность как центральный элемент межперсонального интернет-дискурса / Д. А. Попова // Вестник Бурятского государственного университета. Язык. Литература. Культура. – 2019. – №. 2. – С. 87-91.

56. Попова Е. А. Передача психологических состояний при помощи графических символов в виртуальном пространстве / Е. А. Попова // Сборник материалов XV Международной научной конференции «Психология психических состояний». Казань. – 2021. – С. 435-440.

57. Потебня А. А. Из записок по русской грамматике / А. А. Потебня. – М.: Учпедгиз, 1958. – Т. 1-2. – 536 с.

58. Прайс Дж. Сотрудничество в «невидимом колледже» / Дж. Прайс, Б. Бивер // Коммуникация в современной науке. – М., 1976. – С. 335-350.

59. Преминина М. С. Функционирование окказиональной лексики в интернет-пространстве (на материале социальных сетей) / М. С. Преминина // Культура. Литература. Язык. – 2016. – С. 22-26.

60. Прошкин А. С. Сетевое отчуждение человека в условиях становления информационного общества / А. С. Прошкин // Философия. Политология. – 2019. – С. 42-43.

61. Рыков Ю. Г. Структура и функции онлайн-сообществ: сетевая картография ВИЧ-релевантных групп в социальной сети «ВКонтакте» / Ю. Г. Рыков, О. Ю. Кольцова, П. А. Мейлахс // Социологические исследования. – 2016. – № 8. – С. 30-42.

62. Савва Ю. Б. О проблеме лингвистического анализа сленга в задаче автоматизированного поиска угроз распространения наркомании в виртуальных социальных сетях / Ю. Б. Савва, В. Т. Еременко, Ю. В. Давыдова // Информационные системы и технологии. – 2015. – Т. 92. – №. 6. – С. 68-75.

63. Савельев Д. А. Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации / Д. А. Савельев // Право. Журнал Высшей школы экономики. – 2020. – № 1. – С. 50-74.

64. Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции / С. О. Савчук // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – 2005. – С. 62-88.

65. Сапегина Д. Д. Педагогическая практика в онлайн-режиме: отзывы студентов / Д. Д. Сапегина // Молодежь в меняющемся мире: мировоззренческие основания человека в «текучей современности». – 2020. – С. 19-21.

66. Смелик Н. Д. Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления / Н. Д. Смелик,

А. А. Фильченков // Машинное обучение и анализ данных. – 2016. – Т. 2. – №. 4. – С. 421-441.

67. Смирнова Е. В. Гендерные различия в молодежном Интернет-дискурсе (на материале англоязычных Интернет-блогов и чатов) / Е. В. Смирнова // Лексикографическая копилка: сб. науч. ст. Вып. 8. – СПб.: Изд-во СПбГЭУ, 2019. – С. 73-79.

68. Смирнова О. С. Определение группы риска аккаунтов социальной сети «Вконтакте», попадающих под влияние квестовой игры суицидального характера / О. С. Смирнова // Современные информационные технологии и ИТ-образование. – 2017. – №13 (3). – С. 53-60.

69. Смородина А. А. Интернет-мемы как способ коммуникации человека в современном мире / А. А. Смородина // Международный журнал гуманитарных и естественных наук. – 2019. – №5 (3). – С. 78–82.

70. Стрельников А. И. Исследование методов анализа информационной и лексической насыщенности научных текстов / А. И. Стрельников, М. С. Воробьева // Математическое и информационное моделирование. – 2022. – С. 221-229.

71. Тен Л. В. Тематическое моделирование в задаче автоматической рубрикации новостных текстов / Л. В. Тен // Terra Linguistica. – 2023. – Т. 14. – № 2. – С. 77-91.

72. Тукмакова Н. П. Определение коэффициента взаимной сопряженности в русских и чувашских газетных текстах / Н. П. Тукмакова // Филологические науки. Вопросы теории и практики. – 2020. – Т. 13. – №. 7. – С. 307-312.

73. Тюленева В. Н. Принципы адаптации заимствованной лексики в русском и китайском языках (на примере интернет-обзоров электронной техники) / В. Н. Тюленева // Педагогическое образование в России. – 2016. – №. 11. – С. 100-104.

74. Харари Ф. Теория графов / пер. с англ. В. П. Козырева; под ред. Г. П. Гаврилова. – 2-е изд. – М., 2003. – 300 с.

75. Холодковская Е. В. Особенности синтаксиса англоязычного интернет-комментария социальной сети Facebook / Е. В. Холодковская // Вестник

Волгоградского государственного университета. Серия 2: Языкознание. – 2014. – № 1. – С. 79-83.

76. Хорошевский В. Ф. Семантические технологии в наукометрии: задачи, проблемы, решения и перспективы / В. Ф. Хорошевский, И. Е. Ефименко // Когнитивно-семиотические аспекты моделирования в гуманитарной сфере. Под редакцией В.Л. Стефанюка, Э.А. Тайсиной. – Академия наук РТ, Институт прикладной семиотики АН РТ. – Казань: Изд-во Академии наук РТ, 2017. – С. 222-266.

77. Хохлова М. В. К вопросу о количественном анализе предложно-падежных сочетаний в русском языке на примере законодательных текстов / М. В. Хохлова, В. И. Рубинер // Труды международной конференции «Корпусная лингвистика–2019». – 2019. – С. 149-154.

78. Чеповский А. А. О неявных сообществах на графе взаимодействующих объектов / А. А. Чеповский. // Успехи кибернетики. – 2023. – Т. 4. – № 1. – С. 56-64.

79. Чернявская В. Е. Модусы сетевого пространства: вводные замечания / В. Е. Чернявская // Terra Linguistica. – 2020. – Т. 11. – №. 2. – С. 7-13.

80. Чижик А. В. Исследование динамики общественного настроения в социальных сетях с использованием методов тематического моделирования / А. В. Чижик // International Journal of Open Information Technologies. – 2021. – Т. 9. – № 12. – С. 21-29.

81. Шляховой Д. А. Жанровые характеристики блогов как электронных средств массовой коммуникации / Д. А. Шляховой // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. – 2017. – Т. 8. – №. 4. – С. 939-948.

82. Щурина Ю. В. Интернет-мемы как феномен интернет-коммуникации / Ю. В. Щурина // Научный диалог. – 2011. – №3. – С. 160–172.

83. Юйси М. Языковые средства формирования медиаобраза Китая в русскоязычных интернет-текстах (на примере блогов о китайской опере) / М. Юйси // Филология и человек. – 2021. – №. 1. – С. 169-177.

84. Ярцева В. Н. Лингвистический энциклопедический словарь / В. Н. Ярцева. – М.: Большая российская энциклопедия, 1998. – 685 с.

85. Abaidullah A. M. Identifying Hidden Patterns in Students" Feedback through Cluster Analysis / A. M. Abaidullah, N. Ahmed, E. Ali // International Journal of Computer Theory and Engineering. – 2015. – Vol. 7. – № 1. – P. 16-20.

86. Agarwal S. A topical crawler for uncovering hidden communities of extremist micro-bloggers on Tumblr / S. Agarwal, A. Sureka // 5th workshop on making sense of microposts (MICROPOSTS). – 2015. – P. 26-27.

87. Alba R. D. A graph-theoretic definition of a sociometric clique / R. D. Alba // Journal of Mathematical Sociology. – 1973. – Vol. 3. – № 1. – P. 113-126.

88. Allahyari M. A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling / M. Allahyari, S. Pouriyeh, K. Kochut, H. R. Arabnia // International Journal of Advanced Computer Science and Applications. – Vol. 8. – 2017. – P. 335-349.

89. Al-Maroof R. A. Examining the acceptance of WhatsApp stickers through machine learning algorithms / R. A. Al-Maroof, I. Arpaci, M. Al-Emran, S. A. Salloum, K. Shaalan // Recent advances in intelligent systems and smart applications. – 2021. – P. 209-221.

90. Benko V. Very Large Russian Corpora: New Opportunities and New Challenges / V. Benko, V. Zakharov // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. – 2016. – Iss. 15. – P. 83-98.

91. Bhatia S. Automatic Labelling of Topics with Neural Embeddings / S. Bhatia, J. H. Lau, T. Baldwin // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. – 2016. – P. 953-963.

92. Bianchi F. Cross-lingual contextualized topic models with zero-shot learning / F. Bianchi, S. Terragni, D. Hovy, D. Nozza, E. Fersini // EACL 2021, 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. 2021. – P. 1676-1683.

93. Bindu P. V. Discovering spammer communities in twitter / P. V. Bindu, R. Mishra, P. S. Thilagam // Journal of Intelligent Information Systems. – 2018. – Vol. 51. – P. 503-527.

94. Binesh N. Fuzzy clustering in community detection based on nonnegative matrix factorization with two novel evaluation criteria / N. Binesh, M. Rezghi // Applied Soft Computing. – 2018. – Vol. 69. – P. 689-703.

95. Blei D. Latent Dirichlet Allocation / D. Blei, A. Ng, M. Jordan // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.

96. Bollobas B. Modern Graph Theory / B. Bollobas. – Springer Verlag, New York, USA, 1998. – 397 p.

97. Brunato D. Profiling-UD: a tool for linguistic profiling of texts / D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni // Proceedings of The 12th Language Resources and Evaluation Conference. – 2020. – P. 7145-7151.

98. Chakraborty G. Analysis of unstructured data: Applications of text analytics and sentiment mining / G. Chakraborty, M. Krishna // SAS global forum. – 2014. – P. 1-13.

99. Chakraborty I. Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes / I. Chakraborty, M. Kim, K. Sudhir // Journal of Marketing Research. – 2022. – Vol. 59. – № 3. – P. 600-622.

100. Chaudhary L. Community detection using unsupervised machine learning techniques on COVID-19 dataset / L. Chaudhary, B. Singh // Social Network Analysis and Mining. – 2021. – Vol. 11. – № 1. – P. 1-9.

101. Chen G. B. Word co-occurrence augmented topic model in short text / G. B. Chen, H. Y. Kao // International Journal of Computational Linguistics & Chinese Language Processing. – 2015.– Vol. 20. – № 2. – P. 45-64.

102. Chen P. Y. Deep community detection / P. Y. Chen, A. O. Hero // IEEE Transactions on Signal Processing. – 2015. – Vol. 63. – № 21. – P. 5706-5719.

103. Cheng N. Author gender identification from text / N. Cheng, R. Chandramouli, K. P. Subbalakshmi // Digital investigation. – 2011. – Vol. 8. – № 1. – P. 78-88.

104. Cheng N. Gender identification from e-mails / N. Cheng, X. Chen, R. Chandramouli, K. P. Subbalakshmi // 2009 IEEE Symposium on Computational Intelligence and Data Mining. – 2009. – P. 154-158.

105. Cimino A. Identifying predictive features for textual genre classification: the key role of syntax / A. Cimino, M. Wieling, F. Dell'Orletta, S. Montemagni, G. Venturi // Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it. – 2017. – P. 107-112.

106. Coscia M. A classification for community discovery methods in complex networks / M. Coscia, F. Giannotti, D. Pedreschi // Statistical Analysis and Data Mining: The ASA Data Science Journal. – 2011. – Vol. 4. – № 5. – P. 512-546.

107. Crystal D. Language and the Internet / D. Crystal. – Cambridge: Cambridge University Press. – 2001. – 272 p.

108. Curtotti M. A corpus of Australian Contract Language: Description, profiling and analysis / M. Curtotti, E. C. McCreath // Proceedings of the 13th International Conference on Artificial Intelligence and Law. – 2011. – P. 199-208.

109. Daelemans W. Explanation in computational stylometry / W. Daelemans // International conference on intelligent text processing and computational linguistics. – Springer Berlin Heidelberg, 2013. – P. 451-462.

110. Daelemans W. Overview of PAN 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection / W. Daelemans, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Tschuggnall, M. Wiegmann, E. Zangerle // International conference of the cross-language evaluation forum for european languages. – Springer, Cham, 2019. – P. 402-416.

111. Dell’Orletta F. Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose / F. Dell’Orletta, S. Montemagni, G. Venturi // Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. – 2013. – P. 189-197.

112. Demšar J. Orange: data mining toolbox in Python / J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak,

A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, B. Zupan // The Journal of machine Learning research. – 2013. – Vol. 14. – № 1. – P. 2349-2353.

113. Eder M. Stylometry with R: a package for computational text analysis / M. Eder, J. Rybicki, M. Kestemont // The R Journal. – 2016. – Vol. 8. – № 1 – P. 107-121.

114. Egger R. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts / R. Egger, J. Yu // Frontiers in sociology. – 2022. – Vol. 7. – P. 1-16.

115. Euler L. Solutio problematis ad geometriam situs pertinentis / L. Euler // Commentarii academiae scientiarum Petropolitanae. – 1741. – P. 128-140.

116. Fortunato S. Community detection in graphs / S. Fortunato // Physics reports. – 2010. – Vol. 486. – № 3-5. – P. 75-174.

117. Gmati H. A new algorithm for communities detection in social networks with node attributes / H. Gmati, A. Mouakher, A. Gonzalez-Pardo, D. Camacho // Journal of Ambient Intelligence and Humanized Computing. – 2018. – P. 1-13.

118. Grootendorst M. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics / M. Grootendorst // arXiv preprint arXiv:2203.05794. – 2020. – P. 1-10.

119. Hagen L. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? / L. Hagen // Information Processing & Management. – 2018. – Vol. 54. – №. 6. – P. 1292-1307.

120. Halteren H. V. Linguistic Profiling for Authorship Recognition and Verification / H. V. Halteren // Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). – 2004. – P. 1-8.

121. He K. Hidden community detection in social / K. He, Y. Li, S. Soundarajan, J. E. Hopcroft // Information Sciences. – 2018. – Vol. 425. – P. 92-106.

122. He K. Revealing multiple layers of hidden community structure in networks / K. He, S. Soundarajan, X. Cao, J. Hopcroft, M. Huang // arXiv preprint arXiv:1501.05700. – 2015. – P. 1-10.

123. Hengeveld K. Parts-of-speech systems and morphological types / K. Hengeveld // ACLC Working Papers. – Vol. 2. – 2007. – P. 31-48.

124. Herring S. C. Grammar and electronic communication / S. C. Herring // The encyclopedia of applied linguistics. – 2012. – P. 2338-2346.

125. Hopkins A. Graph theory, social networks and counter terrorism / A. Hopkins. – University of Massachusetts, Darmouth. – 2010. – 22 p.

126. Iqbal F. Wordnet-based criminal networks mining for cybercrime investigation / F. Iqbal, B. C. M. Fung, M. Debbabi, R. Batool, A. Marrington // IEEE Access. – 2019. – Vol. 7. – P. 22740-22755.

127. Jia Y. CommunityGAN: Community Detection with Generative Adversarial Nets / Y. Jia, Q. Zhang, W. Zhang, X. Wang // Proceedings of the World Wide Web. – 2019. – P. 784-794.

128. Jiang J. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel English–Chinese dependency treebank / J. Jiang, H. Liu // Language Sciences. – 2015. – Vol. 50. – P. 93-104.

129. Kamta F. N. A social network analysis of internally displaced communities in northeast Nigeria: potential conflicts with host communities in the Lake Chad region / F. N. Kamta, J. Scheffran // GeoJournal. – 2022. – Vol. 87. – № 5. – P. 4251-4268.

130. Kehoe A. Web corpora / A. Kehoe // A practical handbook of corpus linguistics. – Cham : Springer International Publishing, 2021. – P. 329-351.

131. Kekez M. Model-based imputation of sound level data at thoroughfare using computational intelligence / M. Kekez // Open Engineering. – 2021. – Vol. 11. – № 1. – P. 519-527.

132. Khokhlova M. Big data and word frequency: Measuring the consistency of Russian corpora / M. Khokhlova // Quantitative Approaches to the Russian Language. – Routledge, 2017. – P. 30-48.

133. Kilgariff A. Web as Corpus [Электронный ресурс] / A. Kilgariff, G. Grefenstette. – 2003. – P. 1-15. – URL: https://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf (дата обращения: 25.09.2021).

134. Kirchhoff G. Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanisher Ströme geführt wird / G. Kirchhoff // Annalen der Physik und Chemie. – 1847. – Bd. 72, № 12. – S. 497-508.

135. Köhn A. An Empirical Analysis of the Correlation of Syntax and Prosody / A. Köhn, T. Baumann, O. Dörfler // Proceedings of Interspeech 2018. – 2018. – P. 2157-2161.

136. Kyle K. Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication / K. Kyle. – Ph.D. Thesis, Georgia State University, Atlanta, GA, USA. – 2016. – 186 p.

137. Lafia S. Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network / S. Lafia, L. Fan, A. Thomer, L. Hemphill // Quantitative Science Studies. – 2022. – Vol. 3. – № 3. – P. 694-714.

138. Lenhart A. Social Media & Mobile Internet Use among Teens and Young Adults. Millennials / A. Lenhart, K. Purcell, A. Smith, K. Zickuhr // Pew internet & American life project. – 2010. – P. 1-51.

139. Lilliefors H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown / H. W. Lilliefors // J. Am. Statist. Assoc. – 1967. – Vol. 62. – P. 399-402.

140. Litvinova T. A. Profiling the author of a written text in Russian / T. A. Litvinova // Journal of Language and Literature. – 2014. – Vol. 5. – № 4. – P. 210-216.

141. Litvinova T. Identification of Gender of the Author of a Written Text using Topic-Independent Features / T. Litvinova, P. Seredin, O. Litvinova, O. Zagorovskaya // Pertanika Journal of Social Sciences & Humanities. – 2018. – Vol. 26. – № 1. – P. 103-112.

142. Litvinova T. Profiling the age of Russian bloggers / T. Litvinova, A. Sboev, P. Panicheva // Conference on Artificial Intelligence and Natural Language. – Springer, Cham, 2018. – P. 167-177.

143. Liu H. Dependency distance as a metric of language comprehension difficulty / H. Liu // Journal of Cognitive Science. – 2008. – Vol. 9. – № 2. – P. 159-191.

144. López-Rúa P. Teaching L2 vocabulary through SMS language: some didactic guidelines / P. López-Rúa // ELIA. – 2007. – Vol. 7. – P. 165-188.

145. Lu X. Automatic analysis of syntactic complexity in second language writing / X. Lu // International journal of corpus linguistics. – 2010. – Vol. 15. – № 4. – P. 474-496.

146. Malaterre C. Inferring social networks from unstructured text data: A proof of concept detection of hidden communities of interest / C. Malaterre, F. Lareau // Data & Policy. – 2024. – Vol. 6. – P. 1-19.

147. Mamaev I. Adaptation of Static and Contextualized Topic Modeling Techniques to Hidden Community Detection / I. Mamaev, O. Mitrofanova // International Conference on Internet and Modern Society. – Cham: Springer Nature Switzerland, 2022 (2022b). – P. 85-97.

148. Mamaev I. Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus / I. Mamaev, O. Mitrofanova // Artificial Intelligence and Natural Language. AINL 2020. Communications in Computer and Information Science. – Vol. 1292. – Springer, Cham, 2020 (2020a). – P. 17-33.

149. Mamaev I. D. LiveJournal topic models and their improvement with contextualized representations for creating a model of hidden communities / I. D. Mamaev, O. A. Mitrofanova // International Journal of Open Information Technologies. – 2022 (2022a). – Vol. 10. – № 11. – P. 54-59.

150. Mamaev I. D. The Semantic Shifts of the Topical Structure in the Corpus of Lentach News Posts / I. D. Mamaev, A. A. Mamaeva, D. A. Axenova // Conference on Artificial Intelligence and Natural Language. – Cham : Springer Nature Switzerland, 2022. – P. 27-39.

151. Mamaev I. Hidden Communities in the Russian Social Network Corpus: a Comparative Study of Detection Methods / I. Mamaev, O. Mitrofanova // Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020). – 2020 (2020b). – P. 69-78.

152. Marquardt J. Age and gender identification in social media / J. Marquardt, G. Farnadi, G. Vasudevan, M. F. Moens, S. Davalos, A. Teredesai, M. De Cock // Proceedings of CLEF 2014 Evaluation Labs. – 2014. – Vol. 1180. – P. 1129-1136.

153. McHugh M. L. Interrater reliability: the kappa statistic / M. L. McHugh // Biochemia medica. – 2012. – Vol. 22. – № 3. – P. 276-282.

154. Mehmet M. I. Social media semantics: analyzing meanings in multimodal online conversations / M. I. Mehmet, R. J. Clarke, K. Kautz // International Conference on Information Systems (ICIS) Proceedings, 2014. – P. 1-15.

155. Mei Q. Automatic labeling of multinomial topic models / Q. Mei, X. Shen, C. Zhai // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – 2007. – P. 490-499.

156. Mitrofanova O. E-hypertext Media Topic Model with Automatic Label Assignment / O. Mitrofanova, A. Kriukova, V. Shulginov, V. Shulginov // Recent Trends in Analysis of Images, Social Networks and Texts. – Vol. 1357. – 2021 (2021a). – P. 102-114.

157. Mitrofanova O. Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling / O. Mitrofanova, V. Sampetova, I. Mamaev, A. Moskvina, K. Sukharev // CEUR Workshop Proceedings. – 2021 (2021b). – Vol. 2813. – P. 101-116.

158. Moradi M. Evaluating the robustness of neural language models to input perturbations / M. Moradi, M. Samwald // arXiv preprint arXiv:2108.12237. – 2021. – P. 1-13.

159. Newman M. E. J. The structure and function of complex networks / M. E. J. Newman // SIAM review. – 2003. – Vol. 45. – № 2. – P. 167-256.

160. Nini A. Authorship Profiling in a Forensic Context / A. Nini. – Ph.D. Thesis. Aston University, Birmingham, UK. – 2014. – 250 p.

161. Palese B. Evaluating topic modeling interpretability using topic labeled gold-standard sets / B. Palese, G. Piccoli // Communications of the Association for Information Systems. – 2020. – Vol. 47. – №. 1. – P. 433-451.

162. Paltridge B. Genre analysis and the identification of textual boundaries / B. Paltridge // Applied linguistics. – 1994. – Vol. 15. – № 3. – P. 288-299.

163. Pastor-Satorras R. Evolution and structure of the Internet: A statistical physics approach / R. Pastor-Satorras, A. Vespignani. – Cambridge University Press, 2004. – 267 p.

164. Qi P. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages / P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. – 2020. – P. 101-108.

165. Rainie L. The tone of life on social networking sites / L. Rainie, A. Lenhart, A. Smith // Pew Internet Report. – 2012. – P. 1-30.

166. Salz D. Benavides N., Li J. Hidden Community Detection in Online Forums / D. Salz // CS224W: Machine Learning with Graphs. – 2019. – P. 1-10.

167. Sánchez-Rebollo C. Detection of Jihadism in Social Networks Using Big Data Techniques Supported by Graphs and Fuzzy Clustering / C. Sánchez-Rebollo, C. Puente, R. Palacios, C. Piriz, J. P. Fuentes, J. Jarauta // Complexity. – 2019. – Vol. 2019. – P. 1-13.

168. Scott K. The pragmatics of hashtags: Inference and conversational style on Twitter / K. Scott // Journal of Pragmatics. – 2015. – Vol. 81. – P. 8-20.

169. Sherstinova T. Sentiment Analysis of Literary Texts vs. Reader's Emotional Responses / T. Sherstinova, A. Moskvina, M. Kirina, A. Karysheva, E. Kolpashchikova, P. Maksimenko, A. Seinova, R. Rodionov // 33rd Conference of Open Innovations Association (FRUCT). – IEEE, 2023. – P. 243-249.

170. Skantsi V. Analyzing the unrestricted web: The finnish corpus of online registers / V. Skantsi, V. Laippala // Nordic Journal of Linguistics. – 2023. – P. 1-31.

171. Squires L. Enregistering internet language / L. Squires // Language in Society. – 2010. – 457-492 pp.

172. Tabe C. A. E-morphology in Cameroon social media / C. A. Tabe // US-China Foreign Language. – 2018. – Vol. 16. – № 1. – P. 1-24.

173. Tang Y. (Mis) communication through stickers in online group discussions: A multiple-case study / Y. Tang, K. F. Hew, S. C. Herring, Q. Chen // Discourse & Communication. – 2021. – Vol. 15. – № 5. – P. 582-606.

174. Wang M. Uncovering the Local Hidden Community Structure in Social Networks / M. Wang, B. Li, K. He, J. Hopcroft // ACM Transactions on Knowledge Discovery from Data. – 2023. – Vol. 17. – № 5. – P. 1-25.

175. Wejnert C. An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data / C. Wejnert // Sociological methodology. – 2009. – Vol. 39. – №. 1. – P. 73-116.

176. Wong A. From the hidden to the obvious: classification of primary and secondary school student suicides using cluster analysis / A. Wong, C. C. S. Lai, A. K. Y. Shum, P. S. F. Yip // BMC public health. – 2022. – Vol. 22. – № 1. – P. 1-7.

177. Yang S. Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis / S. Yang, H. Zhang // International Journal of Computer and Information Engineering. – 2018. – Vol. 12. – №. 7. – P. 525-529.

178. Young J. G. Unveiling hidden communities through cascading detection on network structures / J. G. Young, A. Allard, L. Hébert-Dufresne, L. J. Dubé // arXiv preprint arXiv:1211.1364. – 2012. – P. 1-12.

179. Zhou X. Coupling topic modelling in opinion mining for social media analysis / X. Zhou, X. Tao, M. M. Rahman, J. Zhang // Proceedings of the International Conference on Web Intelligence. – 2017. – P. 533-540.