На правах рукописи

УДК:81’33

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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата филологических наук

Санкт-Петербург
2024

Работа выполнена на кафедре математической лингвистики федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет».

Научный руководитель:
Кандидат филологических наук, доцент, доцент кафедры математической лингвистики федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет»
Митрофанова Ольга Александровна

Официальные оппоненты:
доктор филологических наук, профессор, профессор департамента филологии Санкт-Петербургской Школы гуманитарных наук и искусств, заведующий лабораторией языковой конвергенции федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»
Колмогорова Анастасия Владимировна

кандидат филологических наук, доцент кафедры иностранных языков федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»
Клочкова Елена Сергеевна

Ведущая организация:
федеральное государственное автономное образовательное учреждение высшего образования «Южно-Уральский государственный университет (национальный исследовательский университет)»

Защита состоится «**24**» декабря **2024 г.** в **16:00** часов на заседании Совета по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук 33.2.018.08, созданного на базе Российского государственного педагогического университета им. А. И. Герцена, по адресу: 199053, г. Санкт-Петербург, 1-я линия В.О., д. 52, ауд. 48.
С диссертацией можно ознакомиться в фундаментальной библиотеке Российского государственного педагогического университета им. А. И. Герцена (191186, г. Санкт-Петербург, наб. р. Мойки, 48, корп. 5) и на сайте университета по адресу: https://disser.herzen.spb.ru/Preview/Karta/karta_000001049.html

Автореферат разослан « » 2024 г.

Ученый секретарь
диссертационного совета Камшилова Ольга Николаевна

Общая характеристика работы

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 (например, в работах Г.В. Градосельской, N. Binesh, H. Gmati и др.). В криминологии скрытые сетевые сообщества могут служить площадкой для организации незаконных действий, включая киберпреступления, терроризм, торговлю наркотиками и оружием. Изучение таких сообществ помогает правоохранительным органам выявлять угрозы и противостоять им (решению данных задач посвящены исследования Л.О. Кириченко, Н.Л. Аванесян, A. Hopkins и др.). В психологии анализ дискуссий в сетевых сообществах помогает раскрыть многие психологические аспекты, такие как проявление субъективности, поведенческие модели и воздействие на массовое сознание (см. работы А.Н. Воронина, Т.А. Литвиновой, П.В. Паничевой и др.). Наконец, изучение сетевых сообществ может помочь предсказывать и предотвращать социальные конфликты (как показано в исследованиях A.M. Abaidullah, F.N. Kamta, A. Wong). В последнее десятилетие скрытые сообщества попали в сферу интересов компьютерных лингвистов: исследователи подробно изучают коэффициенты лексического разнообразия и логической связности, тематические компоненты групп (труды В.А. Попова, А.А. Чеповского, S. Lafia) 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

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 (см. работы А.С. Прошкина, Е.А. Долгих).

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, инструмента *Profiling-UD*[2], с помощью которого извлекаются основные количественные данные о постах пользователей, а также среда *Microsoft Excel*[3] для проведения статистических расчетов.

Теоретической основой настоящей работы послужили труды отечественных и зарубежных ученых, которые посвящены особенностям функционирования интернет-текстов (Е.В. Холодковская, D. Crystal, L. Squires, S. Herring и др.), определению концепта скрытых сетевых сообществ (П.А. Мейлахс, Ю.Г. Рыков, О.С. Смирнова, А.В. Бухановский и др.), моделированию скрытых сообществ (Д.М. Оболенский, A. Hopkins, M. Wang, A. Wong, L. Chaudhary и др.), описанию разножанровых текстовых коллекций методами компьютерной лингвистики (Н.В. Лукашевич, D. Blei, G.B. Chen и др.).

[1] https://www.python.org/
[2] http://linguistic-profiling.italianlp.it/
[3] 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.

Структура и объем работы. Настоящее диссертационное исследование состоит из введения, трех глав, заключения и списка использованной литературы. Содержание работы представлено на 140 страницах машинописного текста. Список использованных источников насчитывает 179 наименований, из них 84 – на русском языке, 95 – на иностранных языках. Использованные данные представлены в репозитории GitHub[4].

Основное содержание работы

Во **введении** обосновываются тема исследования и ее актуальность, раскрывается степень научной разработанности, определяются объект и предмет, цель и связанные с ней задачи, формулируется гипотеза. Также во введении представлены теоретическая основа и методы исследования, излагаются основные положения, которые выносятся на защиту, указываются рекомендации по

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

использованию предложенной методологии и приводится информация об апробации результатов диссертационного исследования.

В **первой главе диссертации «Веб как корпус: лингвистические особенности функционирования интернет-текстов»** рассматриваются основные параметры интернет-текстов, которые функционируют в русле направления *Web as Corpus*. Так, на графическом уровне одним из востребованных способов выражения эмоций в интернет-текстах являются эмотиконы и стикеры – коммуникационные знаки, которые используются для передачи настроения сообщения и установления доброжелательного контакта. Наравне с вышеупомянутым способом применяется и другое графическое средство – нетрадиционные виды шрифта и комбинации различных цветов, что визуально позволяет выделить текст. Отметим, что на уровне морфосинтаксиса наблюдаются узуальные употребления: нестандартное использование категорий падежа в предложных конструкциях, обильное использование парцелляции и приемов синтаксического сжатия текста, таких как эллипсис и др. Наконец, при изучении интернет-текстов необходимо описать семантические и прагматические особенности постов: внедрение хэштегов как специальной метки принадлежности к тематическому событию, снабжение текста постов специальными устойчивыми концептуальными формами – мемами. Из-за несовершенства алгоритмов автоматической обработки корпусных данных крайне сложно учесть описанные особенности интернет-текстов, поэтому при работе с исследовательским корпусом важно не только описать используемый морфосинтаксический и семантический инструментарий, но и детализировать основные ошибки, которые требуют последующего ручного постредактирования.

Во **второй главе диссертации «Теоретические основания междисциплинарных исследований скрытых сообществ»** систематизируются подходы к определению скрытых сообществ, а также приводится классификация существующих алгоритмов. Отмечается, что в современной терминологии по отношению к термину *‘скрытые сообществ’* существует множество близких понятий. Например, Дж. Прайс вводит термин *‘невидимый колледж’* для описания неформальных связей между учеными в некоторой предметной области. Другим

близким термином является *‘скрытые сообщества по интересам’* (англ. *hidden communities of interest*), вслед за C. Malaterre и F. Lareau этот термин определяется как группы акторов, которые публикуют схожий по семантике медиаконтент, но между которыми не установлены четкие социальные связи. В рамках настоящей работы принимается точка зрения китайской группы исследователей под руководством K. He: скрытые сообщества представляют собой группы пользователей социальных сетей с общими интересами и слабо выраженными или отсутствующими связями. Подобные сетевые структуры сопоставимы с реальными группами с нечеткой организационной структурой (секретные организации и пр.).

Алгоритмы идентификации скрытых сообществ в настоящей работе подразделяются на три группы. К первой группе относятся графовые алгоритмы, которые разрабатываются на основе таких методов, как метод кратчайшего незамкнутого пути, метод поиска ядра графа и др. Во второй группе находятся алгоритмы кластеризации – метода анализа данных, который используется для группировки некоторых объектов с высокой степенью сходства в единые структуры (кластеры). Эти алгоритмы используются зарубежными исследователями Y. Jia, L. Chaudhary, N. Binesh и другими. Наконец, третья группа представлена гибридными алгоритмами, которые, помимо указанных подходов, могут основываться на лингвистических инструментах и базах данных. Так, лингвистическая база данных WordNet используется для выявления преступников в социальных сетях, что представлено к работе арабского исследователя F. Iqbal «Wordnet-Based Criminal Networks Mining for Cybercrime Investigation». В заключении главы акцентируется внимание на том, что алгоритмы обнаружения скрытых сообществ вне зависимости от классификационной группы чувствительны к поставленной исследовательской задаче, экспериментальным данным, доминирующему языку и ряду других факторов.

В **третьей главе диссертации «Разработка процедуры лингвистического профилирования скрытых сообществ»** представлены стадии создания исследовательского алгоритма профилирования. На первом этапе был собран корпус социальных сетей – коллекция русскоязычных текстов из открытого

сегмента социальных сетей, которая подвергается автоматической лингвистической разметке. Для него были сформулированы приведенные ниже положения.

1. Корпус является письменным, так как в него не входят тексты, представленные в аудиоформате (голосовые сообщения, музыкальные композиции и др.).
2. Проводится балансировка корпуса по полу пользователей и временному промежутку (посты, опубликованные не ранее 01.01.2020).
3. Посты для корпуса взяты из одной социальной сети.
4. Выборка данных разделена по пользовательским подкорпусам в соответствии с авторством для упрощения процесса обработки постов и создания тематической модели.
5. Авторы постов не должны быть в категории «Друзья» в социальной сети.

Материалом для создания корпуса стал русскоязычный сегмент социальной сети ВКонтакте объемом более 10 000 постов. Согласно недавнему исследованию аналитической компании Brand Analytics[5], ВКонтакте является лидирующей платформой по количеству сообщений и по количеству активных авторов постов на осень 2023 года. Отметим, что на сегодняшний день разработаны библиотеки для языка программирования Python, которые облегчают процесс выгрузки лингвистической информации и метаинформации с данной платформы.

Лингвистический процесс обработки корпуса был реализован с помощью библиотеки *Stanza*, так как она позволяет непрерывно без переключения на дополнительные модули преобразовать строковую форму представленных данных в леммы с соответствующей морфосинтаксической характеристикой. Также из корпуса были удалены стоп-слова, в нем были размечены лексические конструкции.

Среди главных ошибок автоматической обработки текста, требующих постредактирования, выделим некоторые из них. Во-первых, пользователи

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

используют сочетание букв нескольких алфавитов, имеющих схожее графическое представление (в слове *«восстановление»* некоторые из символов кириллицы заменены на латиницу). При использовании специального сервиса – «Поиск кириллицы в латинице»[6] – установлено, что в слове *«во**сс**тановлени**е**»* присутствует три латинских символа. Слова со схожими комбинациями были добавлены в стоп-словарь. В пользовательских постах встречались и иные комбинации, в частности – кириллица и знаки пунктуации (*« : тарасова»*).

Для создания тематических моделей использовалась многоязычная библиотека *BERTopic*. Для внедрения параметра авторства использовалось деление корпуса по пользователям, организованное на первом этапе эксперимента. Большинство методов тематического моделирования, в том числе BERTopic, не присваивают метки полученным темам. Темой в таком случае является список лемм-тематизаторов. Для удобства интерпретации тем, а в нашем эксперименте – для удобства построения тематических скрытых сообществ, применяется экспертная аннотация тем. Разметка проводилась силами двух экспертов-лингвистов на основании тематической классификации текстов, представленной в Национальном корпусе русского языка, согласованность между ними достигла ≈41% по каппе Коэна. Как отмечает M.L. McHugh, в гуманитарных областях результат согласованности, превышающий 40%, обеспечивает достоверность исследования. Для ручной разметки преодоления порога в 40% также достаточно для того, чтобы в исследовательском наборе данных появились альтернативные метки тем.

С помощью приложений *Easy Linavis* и *Gephi* получен итоговый граф – модель скрытых сообществ. Узлы графа представляют собой пользователей, а связи между ними указывают на наличие тематических пересечений. Густой центр графа указывает, что ряд пользователей публикует посты на большое количество общих тем.

Для настоящей модели в таблице 1 приведена оценка с двух точек зрения: формальной и социально-демографической. Формальные характеристики были

[6] http://invitemsg.com/cyrillic_search.php

получены автоматически в соответствующих режимах *Gephi*, а социально-демографические параметры выгружены из социальной сети ВКонтакте вместе с текстами для исследовательского корпуса.

Таблица 1 — Сводная информация о полученной модели

Параметр	**Значение**
Социодемографические параметры	
Пол	
Мужчины	179
Женщины	197
Год рождения	
(1950, 1952]	1
(1969, 1972]	1
(1972, 1974]	1
(1977, 1980]	3
(1980, 1983]	10
(1983, 1985]	33
(1985, 1988]	68
(1988, 1991]	35
(1991, 1994]	4
(1994, 1996]	1
Город проживания	
Москва	62
Санкт-Петербург	227
Севастополь	0
Города нефедерального значения	56
Информация о высшем образовании	
Указана	120
Не указана	256
Информация об интересах	
Указана	78
Не указана	298
Формальные параметры	
Количество узлов	376
Количество ребер	34507
Плотность графа	0.489
Диаметр графа	3
Тип графа	неориентированный
Средний коэффициент кластеризации графа	0.823
Модулярность	0.167
Предполагаемое количество сообществ на основании расчета модуляции	4

Средний коэффициент кластеризации указывает на большое количество потенциальных групп внутри сети, т.е. она неоднородна. Полученная плотность указывает на средний уровень связанности графа (более 300 узлов). Показатель

модулярности, приближающийся к нулю, позволяет утверждать, что различия в плотностях в группах и между ними явно не выражена. Согласно наблюдениям А.А. Чеповского, итоговые параметры действительно характеризуют структуру социальных сетей, которые обладают рядом особенностей: «маленький диаметр графа (эффект «малого мира»), высокие значения кластерного коэффициента (эффект «транзитивности»)».

Следующий этап описания модели – создание лингвистических профилей скрытых сообществ, т.е. набора языковых признаков, которые выявляются на основе текстовых массивов, составленных участниками той или иной группы. В работе описываются три группы признаков: морфологические корреляции основных частей речи (имя существительное, имя прилагательное, глагол и наречие), синтаксические корреляции средней длины предложения со степенью дистантизации и средней длиной предложных конструкций, а также лексическая корреляция «коэффициент лексической плотности-коэффициент лексического разнообразия».

Среда Profiling-UD, разработанная D. Brunato, стала основным инструментов для сбора количественных показателей. Расчеты проводились в среде Excel, для этого каждый числовой набор анализировался следующим образом.

1. Определение принадлежности количественных параметров выборки нормальному распределению с помощью критерия Колмогорова-Смирнова с учетом дисперсии по формуле (1).

$$D_n^* = D_n(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}) \quad (1)$$

   В формуле (1) 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.

Один из результатов визуализированной зависимости представлен на рисунке 1 (скрытое сообщество «Эзотерика», 17 участников, прямая сила связи $r = 0.4877$ при уровне значимости $p = 0.047$).

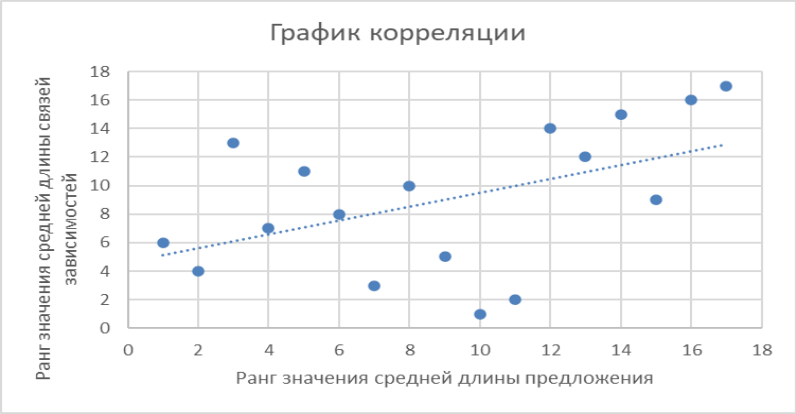

Рисунок 1 — Внутритекстовая корреляция средних значений длины предложений и длины структур зависимостей

В постах пользователя с ID 2644 встречаются примеры инверсионной структуры предложения вида OVS с расширенным дополнением, что приводит к увеличению расстояния между главным и зависимым узлом. На дереве зависимостей, представленной на рисунке 2, степень дистантизации между узлом «является» и левостоящим зависимым дополнением «дверью» равняется трем при 12 узлах синтаксического дерева.

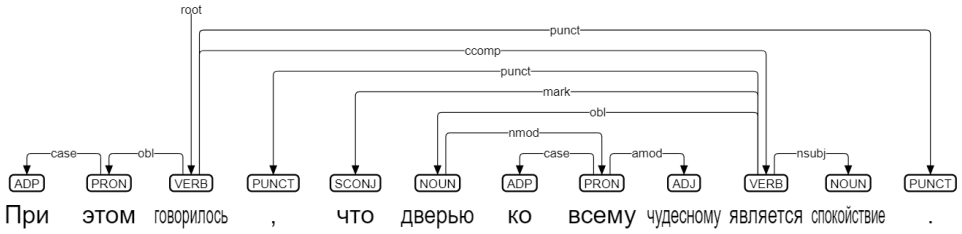

Рисунок 2 — Дерево зависимостей для предложения из поста пользователя с ID 2644

В итоговые лингвистические профили были включены только значимые корреляции. Отметим, что для сообществ численностью менее четырех пользователей расчет корреляций не проводился. Итоги расчетов были сведены воедино (таблица 2), полученные профили были кластеризированы методом Варда. Для каждого объекта кластера рассчитана метрика Silhouette Score, которая показывает силу принадлежности объекта к своему кластеру на основании общности лингвистических признаков.

Таблица 2 — Сводная информация о лингвистических корреляциях в модели скрытых сообществ

Скрытое сообщество	Морфологические корреляции				Синтаксические корреляции		Лексические корреляции
	NV	NAdj	VAdv	AdjAdv	SenLin	LinPrep	TtrDen
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
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	—	—

Итоговый кластер №1 состоит из лингвистических профилей таких скрытых сообществ, как «Эзотерика», «Путешествия», «История», «Здоровье и медицина», «Транспорт», «Природа», «Спорт», «Политика и общественная жизнь», «Досуг, зрелища и развлечения», «Образование», «Строительство и архитектура», «Легкая и пищевая промышленность» и др. Среднее значение метрики Silhouette Score для всех объектов равняется 0.427, что свидетельствует о связности данных внутри кластера. Кластер №2 объединяет лингвистические профили следующих скрытых сообществ: «Частная жизнь», «Дом и домашнее хозяйство», «Бизнес, коммерция, экономика, финансы», «Происшествия» и «Искусство и культура». Среднее значение метрики Silhouette Score чуть выше, чем в кластере №1, – 0.548. Наконец, кластер №3 состоит из лингвистических профилей таких сообществ, как «Наука и технологии», «Армия и государственная безопасность», «Рабочий процесс», «Право», «Религия» и «Психология». Среднее значение метрики Silhouette Score для этого кластера составляет 0.374, что указывает на более низкое качество кластеризации по сравнению с другими кластерами. Отсутствие отрицательных значений Silhouette Score для каждого объекта во всех кластерах указывает на то, что шанс потенциального отнесения лингвистических профилей к другим кластерам низок.

Дисперсионный анализ показал, что наблюдается существенное различие между кластерами скрытых сообществ на синтаксическом уровне. Независимо от исследуемой синтаксической корреляции отмечается наибольшая вариативность для кластеров №1 и №3. Разница между кластерами на уровне морфологии практически несущественна (рисунки 3 и 4).

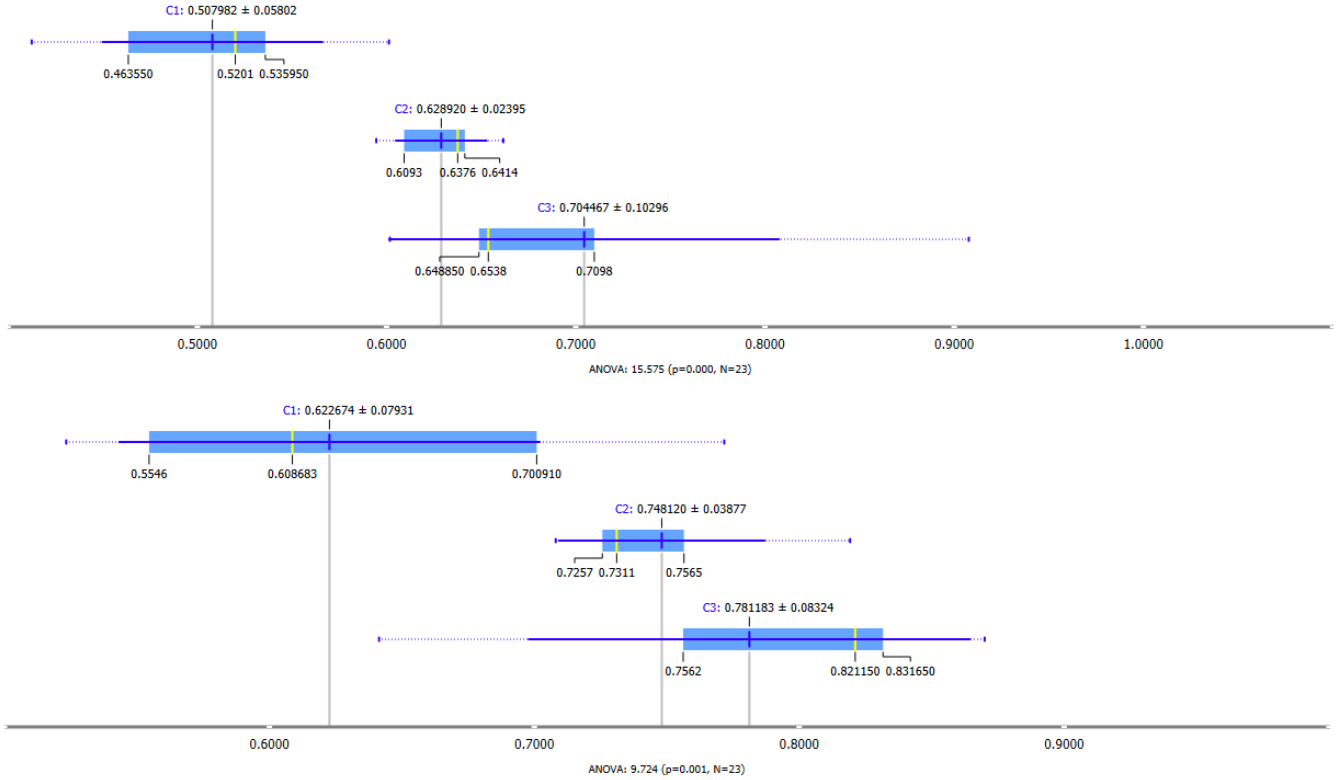

Рисунок 3 — Диаграмма размаха для синтаксических корреляций

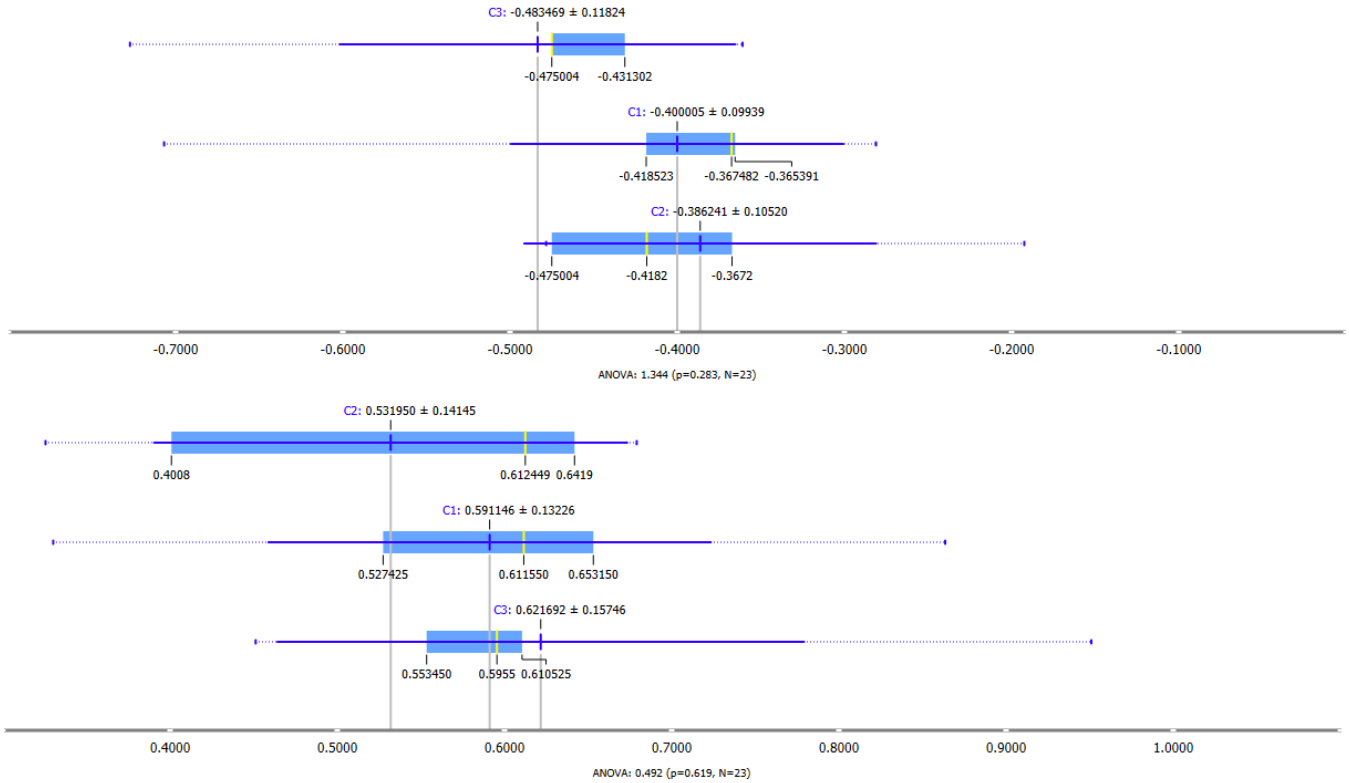

Рисунок 4 — Диаграмма размаха для корреляций «имена существительные-глаголы» и «глаголы-наречия»

Таким образом, в диссертационном исследовании было показано, что предварительно обученные языковые модели BERT и деление исходного корпуса по признаку авторства текстов позволяет настроить алгоритмы тематического моделирования на вывод персонифицированных тематических множеств. «Густой» центр итоговой модели и наличие большого количестве ребер отражает степень скрытой коммуникации пользователей и указывает на способность пользователей создавать политематические публикации. Проверка значимости лингвистических корреляций выявила, что лексические параметры практически не представлены в лингвистических профилях, что затрудняет количественную оценку авторских коммуникативно-письменных навыков. Наконец, результаты иерархической кластеризации с последующим дисперсионным анализом указывают на гетерогенность текстов постов социальных сетей на синтаксическом уровне.

В заключении изложены основные выводы диссертационного исследования, а также приводятся возможные перспективы работы. 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. 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[7].

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

Основные положения исследования отражены в научных докладах, представленных 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).

Публикации по теме диссертационного исследования

Работы по теме диссертационного исследования, опубликованные в рецензируемых научных журналах, входящих в перечень ВАК по научной специальности 5.9.8 общим объемом 2,07 п.л.:

1. **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. (0,96/0,67 п.л.)**
2. **Мамаев И. Д. Лингвистические профили скрытых сообществ: морфосинтаксический аспект / И. Д. Мамаев // Филологические науки. Вопросы теории и практики. – 2024. – Т. 17. – Вып. 4. – С. 1155-1162. (0,57 п.л.)**
3. **Мамаев И. Д. Кластерный анализ лингвистических профилей скрытых сообществ / И. Д. Мамаев // Филологические науки. Вопросы теории и практики. – 2024. – Т. 17. – Вып. 5. – С. 1739-1747. (0,54 п.л.)**

Другие работы, опубликованные по теме диссертационного исследования, общим объемом 2,58 п.л.:

1. 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. – P. 17-33. (0,88/0,61 п.л.)
2. 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. – P. 69-78. (0,5/0,35 п.л.)
3. 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. – P. 85-97. (0,85/0,59 п.л.)
4. 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. – С. 312-315. (0,35 п.л.)