Terra Linguistica. 2024, Том 15, № 1. С. 102–115.
Terra Linguistica. 2024, Vol. 15, No. 1. Pp. 102–115.

Научная статья
УДК 81'32, 81'33
DOI: https://doi.org/10.18721/JHSS.15108

ЛИНГВИСТИЧЕСКИЕ ПАРАМЕТРЫ ДЛЯ ИДЕНТИФИКАЦИИ СКРЫТЫХ СЕТЕВЫХ СООБЩЕСТВ

И.Д. Мамаев[1,2] ✉ , **О.А. Митрофанова**[2]

[1] Балтийский государственный технический университет «Военмех» им. Д.Ф. Устинова, Санкт-Петербург, Российская Федерация;

[2] Санкт-Петербургский государственный университет, Санкт-Петербург, Российская Федерация

✉ mamaev_id@voenmeh.ru

Аннотация. Современные процедуры лингвистической диагностики нуждаются в усовершенствовании применительно к изучению текстов социальных сетей. Одна из проблем, требующих решения, – это выявление лингвистических признаков текстов, значимых для профилирования пользователей – участников скрытых сообществ. Целью данного исследования является разработка гибридного алгоритма обнаружения скрытых сетевых сообществ, учитывающего интересы пользователей, тематику их постов и опирающегося на контекстуализированные языковые модели. Выбор данного подхода обусловлен тем, что алгоритмы выделения скрытых сообществ, основанные на математических методах, используют формальные показатели без учета лингвистических параметров текстов. Это может привести к искажению реального количества и свойств скрытых сообществ. Материалом исследования является корпус русскоязычных постов социальной сети ВКонтакте объемом более 10000 текстов. В результате эксперимента по применению гибридного алгоритма, предложенного авторами статьи, было выделено 34 скрытых сообщества. Авторская методика выявления и профилирования скрытых сообществ представляет интерес для специалистов в области медиаисследований, которые изучают архитектуру социальных сетей. Методику можно внедрить в существующие системы автоматической модерации групп и системы прогнозирования сетевых тенденций.

Ключевые слова: скрытые сообщества, тематическое моделирование, корпусная лингвистика, математические методы, графы.

Для цитирования: Мамаев И.Д., Митрофанова О.А. Лингвистические параметры для идентификации скрытых сетевых сообществ // Terra Linguistica. 2024. Т. 15. № 1. С. 102–115. DOI: 10.18721/JHSS.15108

Research article
DOI: https://doi.org/10.18721/JHSS.15108

LINGUISTIC FEATURES FOR DETECTING HIDDEN NETWORK COMMUNITIES

I.D. Mamaev[1,2] ✉, O.A. Mitrofanova[2]

[1] Baltic State Technical University "Voenmeh" named after D.F. Ustinov, St. Petersburg, Russian Federation;

[2] St. Petersburg State University, St. Petersburg, Russian Federation

✉ mamaev_id@voenmeh.ru

Abstract. Scholars need to improve modern linguistic diagnostic procedures when studying social network texts. One of the unresolved problems is the identification of linguistic features since they are significant for profiling members of hidden communities. The aim of this research is to develop a hybrid algorithm for detecting hidden network communities that takes into account the interests of users, the topics of their posts and is based on contextualized language models. The choice of this approach is due to the fact that algorithms for detecting hidden communities with the help of mathematical methods use formal parameters, but not linguistic ones. This may change the actual number of communities and their properties. The research material is a corpus of VK posts in Russian, which includes more than 10,000 texts. The authors applied the hybrid algorithm and detected 34 hidden communities in the course of the experiment. The current methodology for identifying and profiling hidden communities is of interest to media researchers who study the architecture of social networks. The approach can be implemented into existing automatic group moderation systems and network trend forecasting systems.

Keywords: hidden communities, topic modelling, corpus linguistics, mathematical methods, graphs.

Citation: Mamaev I.D., Mitrofanova O.A., Linguistic features for detecting hidden network communities, Terra Linguistica, 15 (1) (2024) 102–115. DOI: 10.18721/JHSS.15108

Введение

В современном мире наблюдается стремительный рост информационно-коммуникационных технологий, в том числе социальных сетей, которые, с точки зрения компьютерной лингвистики, рассматриваются как корпус текстов, а индивидуальные пользовательские страницы – как ряд подкорпусов, включающие личные тексты и репосты. Исследователи, работая с такими подкорпусами, могут описывать характеристики, на основании которых выявляются скрытые сообщества – группы пользователей, связанные неявными отношениями на основе общности некоторых лингвистических и/или социопсихологических признаков. Изучение скрытых сетевых сообществ имеет важное значение для различных областей знаний. Закрытые форумы и приватные чаты предоставляют социологам уникальную информацию о мнениях общества и менталитете участников коммуникации в чате [1]. В криминологии анализ таких сообществ помогает правоохранительным органам выявлять угрозы и противостоять им [2, 3]. Психологи при анализе сетевых сообществ определяют множество психологических аспектов, таких как характерные поведенческие модели пользователей или способы влияния на массовое сознание [4]. Наконец, изучение сетевых сообществ может предотвратить социальные конфликты [5]. За последнее десятилетие скрытые сообщества привлекли внимание компьютерных лингвистов: исследователи детально изучают коэффициенты лексического разнообразия и логической связности [6, с. 41–42], тематические компоненты групп [7] и другие аспекты. В лингвистике для получения подобного рода результатов зачастую прибегают к системам искусственного интеллекта, что подтверждается исследованиями по фонетике (см., например, [8, 9]), грамматике (см., например, [10, 11]) и другим отраслям. Возросшая потребность в сочетании лингвистических методов и методов

искусственного интеллекта при выявлении и исследовании скрытых сообществ подчеркивает актуальность текущего исследования, целью которого является разработка и практическое применение гибридного алгоритма выявления скрытых сообществ в корпусе социальной сети ВКонтакте. Значимость проводимого исследования определяется распространением корпусов социальных сетей как прикладных наборов данных (датасетов, datasets) для нужд компьютерной лингвистики (например, подкорпус социальных сетей в составе Национального корпуса русского языка (НКРЯ) [12], корпусы, используемые в соревнованиях лингвистических процессоров «Dialogue Evaluation» [13], корпус «Taiga», [14] и т.д.). Предложенный нами алгоритм предполагает внедрение: он совместим с API социальной сети ВКонтакте и способствует улучшению функционала рекомендательной системы данной социальной сети.

Скрытые сообщества и основные алгоритмы их выделения

При определении термина «скрытое сообщество» необходимо отметить, что в научной литературе используется ряд близких понятий. Например, академические социальные сети (Academia.edu, ResearchGate и др.) выступают в качестве площадки для взаимодействия ученых, что позволяет выявить неявные связи между исследователями – «невидимые колледжи». Концепция «невидимых колледжей» была выдвинута Дж. Прайсом в 1960-х гг., однако она подвергалась критике. В частности, Н. Маллинз отмечал, что научная среда организована не как группа с сильными связями, а как некоторая распределенная коммуникативная сеть [15]. В исследовании [16] вводится термин «скрытые сообщества по интересам» (*hidden communities of interest*) – группы деятелей искусств, удовлетворяющие двум требованиям: 1) публикация тематически близкого контента; 2) между участниками группы не установлены явные социальные связи. В рамках настоящего исследования социальные сети рассматриваются как средство повседневного общения, поэтому, расширяя понятия, предложенные в [17, 18], мы определяем скрытые сообщества как сетевые группы пользователей, для которых характерны: 1) общие интересы, 2) отсутствие между двумя пользователями общих «сетевых друзей» как формального показателя знакомства; 3) размытые/отсутствующие социальные связи.

Для алгоритмов обнаружения скрытых сообществ можно выделить три основные архитектуры: графово-математические подходы, кластерный анализ, гибридные подходы. В первой группе используются формальные методы, не учитывающие лингвистические признаки. Например, в [18] представлен двухэтапный алгоритм выявления скрытых сообществ *HICODE* (*HIdden COmmunity DEtection*). Первый этап направлен на определение количества слоев сообществ, что соответствует уровню скрытости изучаемой группы. При обнаружении первого слоя утверждается, что данная группа обладает максимальной степенью связности, и каждый последующий слой имеет меньшую степень связности. Второй этап призван «уточнить» границы более слабых сообществ, так как сообщества с более устойчивыми связями склонны искажать информацию о слабых сообществах. Вторая группа алгоритмов выделения скрытых сообществ основана на разнообразных методах кластерного анализа, таких как иерархическая кластеризация, метод k-средних, метод c-средних и пр. В [19] представлен набор данных о заболеваемости COVID-19 в различных странах. Для выявления групп стран со сходными характеристиками заболеваемости, рассматриваемых как скрытые сообщества, была использована комбинация методов машинного обучения, таких как метод главных компонент для уменьшения размерности данных и метод k-средних для выделения кластеров, которые отображают структуру выделенных скрытых сообществ. Страны, входящие в одно и то же сообщество по результатам анализа, могут взаимодействовать для принятия схожих превентивных мер с целью предотвращения неблагоприятных сценариев развития инфекции. Данные о скрытых сообществах могут использоваться при обосновании решений в области здравоохранения. Наконец, гибридные алгоритмы сочетают несколько формализмов, при этом некоторые могут использовать лингвистические ресурсы. В [20] исследователи

сфокусировались на выявлении скрытого сообщества хакеров с помощью лингвистического тезауруса WordNet. В качестве материала исследования выступали чаты хакеров, из которых вычленялись ключевые слова. Если для ключевых слов двух пользователей был найден общий гипероним в WordNet, то пользователей можно отнести в скрытое сообщество. Подобный подход затрагивает лишь отдельные лексические единицы текстовых массивов. Мы предполагаем, что в гибридных подходах выявления скрытых сообществ необходимо также принимать во внимание синтагматические и парадигматические отношения в текстах, что учитывается в тематических моделях. Поскольку тематическое моделирование является разновидностью нечеткой кластеризации, разрабатываемый нами алгоритм совмещает характеристики первых двух формальных подходов выявления скрытых сообществ и собственно лингвистические параметры текстов, характеризующие их авторов как представителей того или иного скрытого сообщества.

Исследовательская методика

Для создания исследовательского корпуса мы выбрали русскоязычный сегмент социальной сети ВКонтакте, что обусловлено рядом причин. Во-первых, на осень 2023 года данная платформа является ведущей как по количеству сообщений, так и по количеству активных авторов [21]. Во-вторых, для сбора данных нам не требуется предустанавливать виртуальные частные сети. В-третьих, отечественные программисты разработали целый комплекс вспомогательного программного обеспечения для упрощения процедуры выгрузки лингвистической и металингвистической информации, одним из которых является библиотека *vk_api* [22]. При создании корпуса мы придерживались идеи его сбалансированности, под которым понимают «необходимо-достаточное и пропорциональное представление в корпусе текстов различных периодов, жанров, стилей, авторов и т.д., то есть способность отражать все свойства проблемной области» [23, с. 17]. Среди этих параметров мы выделяем следующие.

1. На данный момент фиксация голосовых сообщений в постах сети ВКонтакте не получила широкого распространения, поэтому планируемый корпус будет полностью ***письменным***.

2. С точки зрения пола в корпусе должно быть представлено приблизительно одинаковое количество женщин-пользователей и мужчин-пользователей.

3. Для отражения актуальной статической информации пользователей необходимо ввести фильтр на время публикации текста: не ранее 01.01.2021.

4. Число общих друзей между двумя потенциальными пользователями должно быть строго равно нулю, иначе мы не будем придерживаться определению скрытых сообществ, введенному ранее.

Методом сплошной выборки мы отобрали более 7000 идентификационных номеров пользователей. Далее с помощью специально созданного парсера мы проверяли, являются ли два пользователя явными друзьями в социальной сети. На вход подавался список пользователей, для каждого отдельного пользователя в начале списка проводилась проверка с «хвостом» списка. Если они являлись друзьями, то целевой пользователь удалялся из списка. После проверки в итоговый корпус вошло 10449 постов, написанных 704 пользователями, из них 357 – мужчины, 347 – женщины.

Для создания модели скрытых сообществ на основе лингвистических анализаторов мы предлагаем следующий подход, компоненты которого представлены на рис. 1.

В качестве основного инструмента предобработки мы воспользовались библиотекой *Stanza*. Она включает в себя модули, которые можно последовательно использовать для преобразования строковой формы текста в токены и леммы с соответствующей грамматической информацией [24]. Для минимизации количества неверно распознанных токенов и присвоенных им лемм процесс предобработки нужно повторить несколько раз с последующим анализом. Так, например, в слове «***восс**тановление*» у одного из пользователей при первом визуальном анализе можно не заметить, что часть символов кириллицы заменены на графически похожие латинские символы (см.

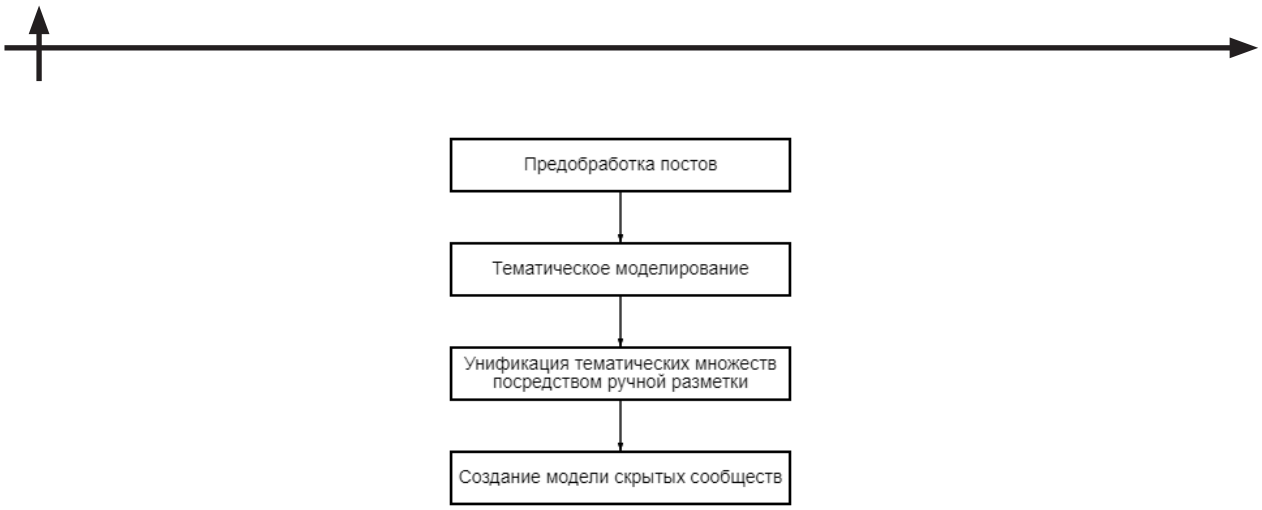

Рис. 1. Лингвистический подход к созданию модели скрытых сообществ
Fig. 1. The linguistic approach to creating the model of hidden communities

выделение). Подобные случаи исправлялись вручную. Далее токены фильтровались с помощью стоп-словаря, в который включены слова широкой семантики: предлоги, союзы, частицы, междометия и т.д. При повторных проверках в словарь добавлялись выявленные нерелевантные единицы: ошибочно распознанные лексемы, обсценная лексика и пр. Наконец, итоговые леммы были дополнены двухсловными сочетаниями с помощью модуля *Phrases* из библиотеки *gensim* [25], так как лексический уровень русского языка представлен не только односложными единицами, но и более длинными конструкциями: «*психологический_помощь*», «*территориальный_зона*», «*главный_архитектор*» и пр.

Для реализации процедур тематического моделирования была выбрана библиотека *BERTopic* [26], поскольку она показывает высокую результативность на корпусах разных жанров, а внедренная в нее контекстуализированная модель *BERT* учитывает полисемию [27, 28]. Сама тематическая модель является мультимодальной, так как, во-первых, многоязычная модель *BERT* позволяет обрабатывать входные тексты, в которых содержатся элементы различных алфавитов, во-вторых, существует возможность интеграции в тематическую модель лексических конструкций, в-третьих, текстовую коллекцию можно разделить на авторские подкорпуса и построить автор-тематическую модель.

В этот алгоритм мы также внедрили дополнительный фильтр. Библиотека *BERTopic* автоматически подбирает количество слов-тематизаторов, которое может существенно варьироваться. Мы выбирали те темы, в которых выводилось 10 лемм, поскольку стандартные библиотеки, работающие с LDA-моделями и их модификациями, ориентированы именно на это количество лемм [29, 30]. Таким образом, с учетом описанных выше особенностей для 704 пользователей только 376 авторам были присвоены тематические характеристики (53% от общего объема корпуса), что связано с рядом причин. 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 нерепрезантативной из-за пересечения множеств слов. Во-вторых, стандартные процедуры фильтрации текстов стоп-словарем уменьшают объём корпуса за счет удаления «шумных» единиц, в результате чего обработанные пользовательские подкорпуса с малым количеством лексем не обрабатываются алгоритмом. Если не включать в стандартные процедуры предобработки текстов фильтрацию стоп-словарем, то итоговые наборы тем будет сложно интерпретировать лингвистам-экспертам. Примеры сгенерированных контекстуализированной моделью наборов тем приведены в табл. 1.

Таблица 1. Наборы тем для отдельных пользователей
Table 1. Topical sets for particular users

ID пользователя	Тема	Слова-тематизаторы
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
33970	0	флот, фрегат, морской, ракета, ракетный, адмирал, горшков, цель, балтийский, море
	1	завод, оборудование, позволить, кремний, российский, тонна, несколько, насос, скважина, создать
		сердце, взгляд, судьба, встреча, суббота, слышать, безнадежный, начаться, подобный, погнать

Полученные в результате исследования темы требуют обобщения с помощью меток, назначаемых в процессе разметки. Тематическая разметка проводилась на основе текстовой классификации, представленной в Национальном корпусе русского языка (НКРЯ) [31], которую мы незначительно расширили в зависимости от анализируемых лемм-тематизаторов. Наряду с такими метками, как «бизнес, коммерция, экономика, финансы»; «дом и домашнее хозяйство»; «здоровье и медицина»; «зрелища и развлечения»; «искусство»; «криминал»; «наука (по разделам и отраслям)»; «политика и общественная жизнь»; «право»; «производство»; «сельское хозяйство»; «спорт»; «природа»; «частная жизнь» и т.д., мы ввели дополнительные метки: «журналистика», которая описывает труд работников прессы, а также «рабочий процесс», в рамках которой представлен непосредственный процесс создания чего-либо. Оценка тем производилась силами двух лингвистов-экспертов. Они должны были ознакомиться со тематическими множествами слов и предлагаемой для них меткой; при согласии с предложенной темой они ставили 1, а при несогласии — 0. Итоги опроса были представлены в виде матрицы согласованности (табл. 2).

Таблица 2. Матрица для расчета каппы Коэна
Table 2. Matrix for calculating Cohen's kappa

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

В результате вычислений в среде Excel итоговое значение каппы Коэна равняется 40.86%, что удовлетворяет минимальным требованиям при проведении оценки согласованности в социогуманитарных исследованиях, согласно работам [32, 33].

Для построения модели скрытых сообществ мы можем воспользоваться комбинацией двух графовых инструментов – *Easy Linavis* [34] и *Gephi* [35]. Узлами графа являются идентификационные номера пользователей, а ребрами – скрытая связь (рис. 2).

Результаты и обсуждение

В результате применения предложенного нами гибридного алгоритма было выявлено 34 скрытых сообщества (рис. 3), охватывающие как узконаправленные, так и широкопрофильные

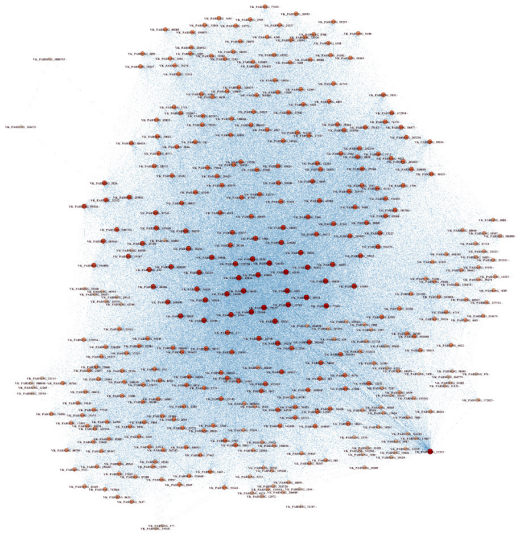

Рис. 2. Модель скрытых сообществ социальной сети ВКонтакте
Fig. 2. Model of hidden communities of VK social network

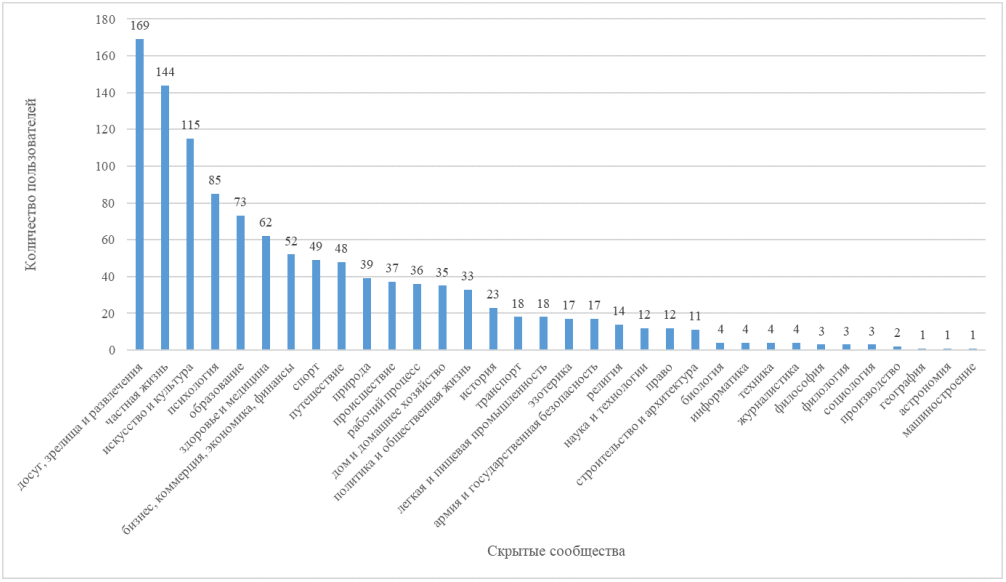

Рис. 3. Количество пользователей в скрытых сообществах
Fig. 3. The number of users in hidden communities

тематики. Узконаправленные тематики представляют интерес для определенных социальных и/или профессиональных групп (например, философы, журналисты, филологи и т.д.), широкопрофильные тематики затрагивают глобальные вопросы, для которых нельзя однозначно определить социальную и/или профессиональную группу (например, сообщества «Частная жизнь», «Природа»).

На рис. 4 представлены данные по количеству сообществ, в которых состоит каждый отдельный пользователь (всего пользователей – 376). «Хвост» полученного распределения указывает на приверженность единому интересу, пользователи указывают в среднем от двух до четырех интересов.

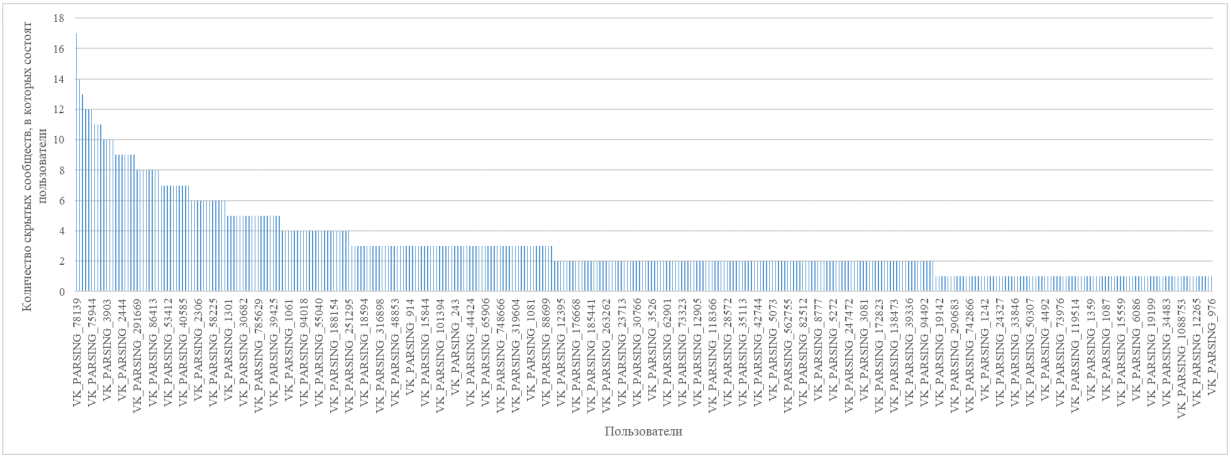

Рис. 4. Количество скрытых сообществ на одного пользователя

Fig. 4. Number of hidden communities per user

По данным о тематике постов можно сформировать лингвистический профиль пользователей, посты которых вошли в исследовательский корпус: это лица, интересы которых сосредоточены на ограниченном числе тем (от двух до четырех), при этом самыми популярными для обсуждения являются три темы: «досуг, зрелища и развлечения», «частная жизнь», искусство и культура». К ним примыкают темы «психология», «образование», «здоровье и медицина», «бизнес, коммерция, экономика, финансы», «спорт», «путешествие», «природа» и т.д. Полученные группы пользователей могут использоваться разработчиками ВКонтакте для совершенствования системы рекомендаций, функционирующей внутри социальной сети.

Для нашего графа мы можем дать оценку с двух точек зрения: формальной и социально-демографической. Формальные характеристики были получены автоматически при расчетах в соответствующих режимах *Gephi*.

Таблица 3. Формальные характеристики графа
Table 3. Formal features of the graph

Параметр	Значение
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

Для автоматически созданного графа существует два подхода к оценке 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. 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

Таблица 4. Социально-демографические параметры модели
Table 4. Social and demographic parameters of the model

Параметр	Значение
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

группы. Сейчас мы будем исходить из предположения, что реальное количество сообществ в модели нельзя вычислить.

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, близкий к 0, позволяет утверждать, что различия в плотностях в группах и между ними пределами явно не выражена. Согласно [36], полученные параметры действительно характеризуют структуру социальных сетей, которые имеют ряд особенностей: «маленький диаметр графа (эффект «малого мира»), высокие значения кластерного коэффициента (эффект «транзитивности»)...».

Также с помощью библиотеки *vk_api* мы извлекли базовые социально-демографические параметры пользователей скрытых сообществ. Характеристики представлены в табл. 4. На основе полученных данных можно составить социологический профиль пользователей, посты которых вошли в корпус: это лица среднего возраста, проживающие преимущественно в Санкт-Петербурге. Несмотря на то, что некоторые пользователи не попали в итоговую модель, число участников

сообществ мужского и женского пола оказалось приблизительно равным. Возрастные данные подтверждают результаты исследований [37, 38]: активными пользователями социальных сетей являются люди в возрасте от 30 до 49 лет. Отметим, что ряд параметров (образование и интересы), пользователи предпочли не оставлять в открытом доступе.

При подсчете данных о возрасте мы не включили информацию о трех пользователях, хотя они и указали полную дату рождения, поскольку она не соотносилась с визуальной информацией, которая представлена на странице. К ним относились, например, фотографии и картинки, которые не характерны для данного возрастного периода. Эти пользователи указали следующие даты: 11.11.1918, 26.11.1925 и 11.11.1928. В частности, при анализе сетевых данных пользователя 1925-го года рождения оказалось, что он публикует посты, один из которых является благодарностью за поздравление с прошедшим днем рождения. К тексту прикреплен аудиофайл с песней группы «Сектор Газа» – «Мне снова 30 лет»[1].

Данные о скрытых сообществах пользователей социальной сети ВКонтакте, полученные в ходе применения разработанного нами алгоритма, позволяют сформировать для каждого сообщества социолингвистический профиль, учитывающий те же параметры говорящих, которые используются при исследовании авторского идиостиля и социально-психологических характеристик участников сетевой коммуникации [39].

Заключение

Результаты исследования, представленные в статье, подтверждают применимость предложенного нами алгоритма выявления скрытых сообществ пользователей социальных сетей. Разработанный алгоритм выявления скрытых сообществ является гибридным: он позволяет строить графовое представление структуры скрытых сообществ, при этом основан на алгоритмах кластеризации текстовых данных и учитывает лингвистические параметры постов пользователей, прежде всего, их тематику. Среди особенностей нашего подхода следует указать использование мультимодального тематического моделирования с привлечением контекстуализированной языковой модели BERT с учетом авторства постов. Данные условия экспериментов обеспечивают настройку алгоритма выявления скрытых сообществ на выделение узконаправленных персонифицированных тематических списков.

Дальнейшие исследования связаны с многоуровневым лингвистическим анализом текстов пользователей – участников скрытых сообществ. Предполагается изучение внутритекстовых корреляций на трех уровнях – морфологическом, синтаксическом и лексическом, что позволит составить расширенный лингвистический профиль пользователей скрытых сообществ.

СПИСОК ИСТОЧНИКОВ

1. **Градосельская Г.В., Щеглова Т.Е., Карпов И.А.** Картирование политически активных групп в Фейсбуке: динамика 2013-2018 гг. // Вопросы кибербезопасности. 2019. № 4 (32). С. 94–104. DOI: 10.21681/2311-3456-2019-4-94-104

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

3. **Аванесян Н.Л. и др.** Выявление значимых признаков противоправных текстов // Вопросы кибербезопасности. 2020. № 4 (38). С. 76–84. DOI: 10.21681/2311-3456-2020-04-76-84

4. **Воронин А.Н., Ковалева Ю.В., Чеповский А.А.** Взаимосвязь сетевых характеристик и субъектности сетевых сообществ в социальной сети Твиттер // Вопросы кибербезопасности. 2020. № 3 (37). С. 40–57. DOI: 10.21681/2311-3456-2020-03-40-57

[1] Пост пользователя VK с id 1088753 [Электронный ресурс]. URL: https://vk.com/id1088753?w=wall1088753_26908 (дата обращения: 08.08.2023).

5. **Kamta F.N., Scheffran J.** A social network analysis of internally displaced communities in northeast Nigeria: potential conflicts with host communities in the Lake Chad region // GeoJournal. 2022. Vol. 87. No. 5. Pp. 4251–4268. DOI: 10.1007/s10708-021-10500-8

6. **Попов В.А., Чеповский А.А.** Выделение неявных пересекающихся сообществ на графе взаимодействия Telegram-каналов с помощью «метода Галактик» // Труды института системного анализа российской академии наук. 2022. Т. 72. № 4. С. 39–50. DOI: 10.14357/20790279220405

7. **Lafia S. et al.** Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network // Quantitative Science Studies. 2022. Vol. 3. No. 3. Pp. 694–714. DOI: 10.1162/qss_a_00209

8. **Zimina S., Evdokimova V.** Acoustic Characteristics of Speech Entrainment in Dialogues in Similar Phonetic Sequences // Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23. Springer International Publishing, 2021. Pp. 818–825. DOI: 10.1007/978-3-030-87802-3_73

9. **Zhou X. et al.** Coupling topic modelling in opinion mining for social media analysis // Proceedings of the International Conference on Web Intelligence. 2017. Pp. 533–540. DOI: 10.1145/3106426.3106459

10. **Bell S., Yannakoudakis H., Rei M.** Context is key: Grammatical error detection with contextual word representations // arXiv preprint arXiv:1906.06593. 2019. Pp. 1–13. DOI: 10.48550/arXiv.1906.06593

11. **He Z.** English grammar error detection using recurrent neural networks // Scientific Programming. 2021. Vol. 2021. Pp. 1–8. DOI: 10.1155/2021/7058723

12. Корпус социальных сетей. URL: https://ruscorpora.ru/search?search=CgQyAggWMAE%3D (дата обращения: 08.08.2023).

13. Dialogue Evaluation. URL: https://www.dialog-21.ru/evaluation/ (дата обращения: 08.08.2023).

14. Taiga Corpus. URL: https://tatianashavrina.github.io/taiga_site/ (дата обращения: 08.08.2023).

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

16. **Malaterre C., Lareau F.** Inferring social networks from unstructured text data: A proof of concept detection of hidden communities of interest // Data & Policy. 2024. Vol. 6. Pp. 1–19. DOI: 10.1017/dap.2023.48

17. **Fortunato S.** Community detection in graphs // Physics reports. 2010. Vol. 486. No. 3-5. Pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002

18. **He K. et al.** Hidden community detection in social // Information Sciences. 2018. Vol. 425. Pp. 92–106. DOI: 10.1016/j.ins.2017.10.019

19. **Chaudhary L., Singh B.** Community detection using unsupervised machine learning techniques on COVID-19 dataset // Social Network Analysis and Mining. 2021. Vol. 11. No. 1. Pp. 1–9. DOI: 10.1007/s13278-021-00734-2

20. **Iqbal F. et al.** Wordnet-based criminal networks mining for cybercrime investigation // IEEE Access. 2019. Vol. 7. Pp. 22740–22755. DOI: 10.1109/ACCESS.2019.2891694

21. Социальные сети в России: цифры и тренды, осень 2023. URL: https://brandanalytics.ru/blog/social-media-russia-autumn-2023/ (дата обращения: 02.02.2024).

22. Документация vk_api. URL: https://vk-api.readthedocs.io/en/latest/ (дата обращения: 07.08.2023).

23. **Захаров В.П., Богданова С.Ю.** Корпусная лингвистика: учебник для студентов гуманитарных вузов. Иркутск: ИГЛУ, 2011. 161 с.

24. **Qi P. et al.** Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020. Pp. 101–108. DOI: 10.48550/arXiv.2003.07082

25. Gensim. URL: https://radimrehurek.com/gensim/ (дата обращения: 07.08.2023).

26. **Grootendorst M.** BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics // arXiv preprint arXiv:2203.05794. 2020. Pp. 1–10. DOI: 10.48550/arXiv.2203.05794

27. **Egger R., Yu J.** A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts // Frontiers in sociology. 2022. Vol. 7. Pp. 1–16. DOI: 10.3389/fsoc.2022.886498

28. **Митрофанова О.А., Атугодаге М.М.** Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. 2023. Т. 14. № 1. С. 70–87. DOI: 10.18721/JHSS.14107

29. **Gan J., Qi Y.** Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example // Entropy. Vol. 23 (10). 2021. Pp. 1–45. DOI: 10.3390/e23101301

30. **Hasan M. et al.** Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA) // Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Springer, Singapore. 2021. Pp. 341–354. DOI: 10.1007/978-981-33-4673-4_27

31. Параметры текстов. URL: https://ruscorpora.ru/page/instruction-parameter/ (дата обращения: 08.08.2023).

32. **McHugh M.L.** Interrater reliability: the kappa statistic // Biochemia medica. 2012. Vol. 22. No. 3. Pp. 276–282.

33. Probability and Statistics Topics Index. URL: https://www.statisticshowto.com/probability-and-statistics/ (дата обращения: 08.08.2023).

34. Easy Linavis. URL: https://ezlinavis.dracor.org/ (дата обращения: 08.08.2023).

35. The Open Graph Viz Platform. URL: https://gephi.org/ (дата обращения: 08.08.2023).

36. **Чеповский А.А.** О неявных сообществах на графе взаимодействующих объектов // Успехи кибернетики. 2023. Т. 4. № 1. С. 56–64. DOI: 10.51790/2712-9942-2023-4-1-08

37. **Lenhart A. et al.** Social Media & Mobile Internet Use among Teens and Young Adults. Millennials // Pew internet & American life project. 2010. Pp. 1–51.

38. **Rainie L., Lenhart A., Smith A.** The tone of life on social networking sites // Pew Internet Report. 2012. Pp. 1–30.

39. **Литвинова Т.А.** Стилеметрическая идентификация автора текста. Истоки, Воронеж, 2022. 254 с.

REFERENCES

[1] **Gradoselskaya G.V., Scheglova T.E., Karpov I.** Mapping of politically active groups on Facebook: dynamics of 2013–2018, Voprosy kiberbezopasnosti [Issues of cybersecurity]. 4 (32) (2019) 94–104. DOI: 10.21681/2311-3456-2019-4-94-104

[2] **Kirichenko L.O., Radivilova T.A., Baranovsky A.** Obnaruzheniye kiberugroz s pomoshchyu analiza sotsialnykh setey [Detecting cyberthreats with the help of analyzing social networks], International Journal "Information Technologies & Knowledge". 11 (1) (2017) 23–48.

[3] **Avanesyan N.L. et al.** Identifying the significant features in illegal texts, Voprosy kiberbezopasnosti [Issues of cybersecurity]. 4 (38) (2020) 76–84. DOI: 10.21681/2311-3456-2020-04-76-84

[4] **Voronin A.N., Kovaleva Yu.V., Chepovskiy A.A.** Interconnection of network characteristics and subjectivity of network communities in the social network Twitter, Voprosy kiberbezopasnosti [Issues of cybersecurity]. 3 (37) (2020) 40–57. DOI: 10.21681/2311-3456-2020-03-40-57

[5] **Kamta F.N., Scheffran J.** A social network analysis of internally displaced communities in northeast Nigeria: potential conflicts with host communities in the Lake Chad region, GeoJournal. 87 (5) (2022) 4251–4268. DOI: 10.1007/s10708-021-10500-8

[6] **Popov V.A., Chepovskiy A.A.** Use of the "Galaxies method" to reveal overlapping communities on the Telegram channels interraction graph, Proceedings of the institute for systems analysis of the Russian academy of sciences. 72 (4) 2022 39–50. DOI: 10.14357/20790279220405

[7] **Lafia S. et al.** Subdivisions and crossroads: Identifying hidden community structures in a data archive's citation network, Quantitative Science Studies. 3 (3) (2022) 694–714. DOI: 10.1162/qss_a_00209

[8] **Zimina S., Evdokimova V.** Acoustic Characteristics of Speech Entrainment in Dialogues in Similar Phonetic Sequences, Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23. Springer International Publishing, 2021. Pp. 818–825. DOI: 10.1007/978-3-030-87802-3_73

[9] **Zhou X. et al.** Coupling topic modelling in opinion mining for social media analysis, Proceedings of the International Conference on Web Intelligence. (2017) 533–540. DOI: 10.1145/3106426.3106459

[10] **Bell S., Yannakoudakis H., Rei M.** Context is key: Grammatical error detection with contextual word representations, arXiv preprint arXiv:1906.06593. (2019) 1–13. DOI: 10.48550/arXiv.1906.06593

[11] **He Z.** English grammar error detection using recurrent neural networks, Scientific Programming. (2021) 1–8. DOI: 10.1155/2021/7058723

[12] Korpus sotsialnykh setey [Social Network Corpus]. Available at: https://ruscorpora.ru/search?-search=CgQyAggWMAE%3D (accessed 08.08.2023).

[13] Dialogue Evaluation. Available at: https://www.dialog-21.ru/evaluation/ (accessed 08.08.2023).

[14] Taiga Corpus. Available at: https://tatianashavrina.github.io/taiga_site/ (accessed 08.08.2023).

[15] **Mullins N.C.** Analiz soderzhaniya neformalnoy kommunikatsii mezhdu biologami [Analysis of the content of informal communication among biologists], Kommunikatsiya v sovremennoy nauke [Communication in modern science]. (1976) 239–260.

[16] **Malaterre C., Lareau F.** Inferring social networks from unstructured text data: A proof of concept detection of hidden communities of interest, Data & Policy. 6 (2024) 1–19. DOI: 10.1017/dap.2023.48

[17] **Fortunato S.** Community detection in graphs, Physics reports. 486 (3-5) 2010 75–174. DOI: 10.1016/j.physrep.2009.11.002

[18] **He K. et al.** Hidden community detection in social // Information Sciences. 425 (2018) 92–106. DOI: 10.1016/j.ins.2017.10.019

[19] **Chaudhary L., Singh B.** Community detection using unsupervised machine learning techniques on COVID-19 dataset, Social Network Analysis and Mining. 11 (1) (2021) 1–9. DOI: 10.1007/s13278-021-00734-2

[20] **Iqbal F. et al.** Wordnet-based criminal networks mining for cybercrime investigation, IEEE Access. 7 (2019) 22740–22755. DOI: 10.1109/ACCESS.2019.2891694

[21] Sotsialnyye seti v Rossii: tsifry i trendy, osen 2023 [Social networks in Russia: figures and trends, autumn 2023]. Available at: https://brandanalytics.ru/blog/social-media-russia-autumn-2023/ (accessed 02.02.2024).

[22] Dokumentatsiya vk_api [vk_api documentation]. Available at: https://vk-api.readthedocs.io/en/latest/ (accessed 07.08.2023).

[23] **Zakharov V.P., Bogdanova S.Yu.** Korpusnaya lingvistika: uchebnik dlya studentov gumanitarnykh vuzov [Corpus linguistics: a textbook for students of humanitarian universities]. Irkutsk: IGLU, 2011. 161 p.

[24] **Qi P. et al.** Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. (2020) 101–108. DOI: 10.48550/arXiv.2003.07082

[25] Gensim. Available at: https://radimrehurek.com/gensim/ (accessed 07.08.2023).

[26] Grootendorst M., BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics, arXiv preprint arXiv:2203.05794. (2020) 1–10. DOI: 10.48550/arXiv.2203.05794

[27] **Egger R., Yu J.** A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts, Frontiers in sociology. 7 (2022) 1–16. DOI: 10.3389/fsoc.2022.886498

[28] **Mitrofanova O.A., Athugodage M.M.** Dynamic topic modelling of the Russian legal text corpus, Terra Linguistica. 14 (1) (2023) 70–87. DOI: 10.18721/JHSS.14107

[29] **Gan J., Qi Y.** Selection of the Optimal Number of Topics for LDA Topic Model–Taking Patent Policy Analysis as an Example, Entropy. 23 (10) (2021) 1–45. DOI: 10.3390/e23101301

[30] **Hasan M. et al.** Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA), Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Springer, Singapore, 2021. Pp. 341–354. DOI: 10.1007/978-981-33-4673-4_27

[31] Parametry tekstov [Text Parameters]. Available at: https://ruscorpora.ru/page/instruction-parameter/ (accessed 08.08.2023).

[32] **McHugh M.L.** Interrater reliability: the kappa statistic, Biochemia medica. 22 (3) (2012) 276–282.

[33] Probability and Statistics Topics Index. Available at: https://www.statisticshowto.com/probability-and-statistics/ (accessed 08.08.2023).

[34] Easy Linavis. Available at: https://ezlinavis.dracor.org/ (accessed 08.08.2023).

[35] The Open Graph Viz Platform. Available at: https://gephi.org/ (accessed 08.08.2023).

[36] **Chepovskiy A.A.** Implicit Communities Defined on the Graph for Interacting Objects, Russian Journal of Cybernetics. 4 (1) (2023) 56–64 DOI: 10.51790/2712-9942-2023-4-1-08

[37] **Lenhart A. et al.** Social Media & Mobile Internet Use among Teens and Young Adults. Millennials, Pew internet & American life project. (2010) 1–51.

[38] **Rainie L., Lenhart A., Smith A.** The tone of life on social networking sites, Pew Internet Report. (2012) 1–30.

[39] **Litvinova T.A.** Stilemetricheskaya identifikatsiya avtora teksta [Stylometric identification of the author of the text]. Istoki, Voronezh, 2022.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Мамаев Иван Дмитриевич
Ivan D. Mamaev
E-mail: mamaev_id@voenmeh.ru
https://orcid.org/0000-0003-3362-9131

Митрофанова Ольга Александровна
Olga A. Mitrofanova
E-mail: o.mitrofanova@spbu.ru
https://orcid.org/0000-0002-3008-5514

Поступила: 20.02.2024; Одобрена: 04.03.2024; Принята: 07.03.2024.
Submitted: 20.02.2024; Approved: 04.03.2024; Accepted: 07.03.2024.