DOI: 10.24412/2070-1381-2023-97-70-84

Разработка системы лингвистических маркеров для автоматизированной выгрузки тематических текстовых данных из социальной сети[1]

Саркисова Анна Юрьевна
Кандидат филологических наук, доцент, научный сотрудник, факультет государственного управления, МГУ имени М.В. Ломоносова, Москва, РФ.
E-mail: sarkisova@data.tsu.ru
SPIN-код РИНЦ: 1212-0879
ORCID ID: 0000-0001-5674-0962

Петров Евгений Юрьевич
Техник, суперкомпьютерный центр, Национальный исследовательский Томский государственный университет, Томск, РФ.
E-mail: petrov@data.tsu.ru
SPIN-код РИНЦ: 6469-0644
ORCID ID: 0000-0002-7140-7882

Дунаева Дарья Олеговна
Научный сотрудник, факультет государственного управления, МГУ имени М.В. Ломоносова, Москва, РФ.
E-mail: ddo@data.tsu.ru
SPIN-код РИНЦ: 7164-7368
ORCID ID: 0000-0002-6622-9882

Аннотация
Автоматизированный поиск и отбор текстов по определенной теме в целевом источнике для формирования репрезентативной тематической текстовой коллекции (текстового датасета) большой размерности, будучи частным случаем получения и структурирования первичных данных, остается одной из наиболее востребованных прикладных задач обработки естественного языка. В статье представлен опыт разработки системы лингвистических маркеров, позволяющей извлекать автоматизированными методами тексты, связанные с тематикой вакцинации от COVID-19, на материале социальной сети «ВКонтакте». Для формирования итогового датасета использовалась комбинация лингвистических методов с методами сбора и обработки текстовых данных. Тестовый список маркеров сформирован на основе фоновых знаний, работы со словарями и специальными лингвистическими сервисами. Ставилась задача сформировать список слов, объединенных общим концептуальным признаком, спрогнозировать совместную встречаемость слов в текстах о вакцинации от COVID-19 или найти специфичные слова, маркирующие данную тему: окказионализмы, обозначения специфичных реалий. Контент выгруженных с помощью тестового списка маркеров тематических сообществ в сети «ВКонтакте» стал источником автоматизированного и экспертного извлечения основного массива маркеров (354 единицы). Подробно описана процедура автоматизированной фильтрации промежуточной текстовой выборки (12,8 млн текстов); приведена методика формирования стоп-слов. За период с 01.01.2020 по 01.03.2023 извлечено 4,5 млн релевантных сообщений; валидность маркеров подтвердилась незначительным в масштабе больших данных количеством шума. Систематизированы общие принципы подготовки лингвистических маркеров для автоматизированной выгрузки больших текстовых данных; отмечены сильные и слабые стороны данного инструмента; предложены рекомендации по формированию списка лингвистических маркеров.

Ключевые слова
Лингвистический маркер, большие данные, автоматизированный сбор данных, выгрузка данных, текстовая коллекция, полнотекстовый поиск, социальные сети, «ВКонтакте», открытый API, вакцинация, COVID-19.

Development of a System of Linguistic Markers for Automated Unloading of Thematic Text Data from a Social Network[2]

Anna Yu. Sarkisova
PhD, Associate Professor, Research Associate, School of Public Administration, Lomonosov Moscow State University, Moscow, Russian Federation.
E-mail: sarkisova@data.tsu.ru
ORCID ID: 0000-0001-5674-0962

Evgeny Yu. Petrov
Technician, Supercomputer Center, National Research Tomsk State University, Tomsk, Russian Federation.
E-mail: petrov@data.tsu.ru
ORCID ID: 0000-0002-7140-7882

Daria O. Dunaeva
Research Associate, School of Public Administration, Lomonosov Moscow State University, Moscow, Russian Federation.
E-mail: ddo@data.tsu.ru
ORCID ID: 0000-0002-6622-9882

[1] Исследование выполнено при финансовой поддержке РНФ, проект 23-28-01025 «Исследование нарративов в социальных медиа с применением технологии анализа больших данных (на примере нарративов о вакцинации от COVID-19)».
[2] The study was carried out with the financial support of the RSCF, project 23-28-01025 “Research of narratives in social media using big data analysis technology (using the example of narratives about vaccination against COVID-19)”.

Abstract
Automated search and selection of texts on a specific topic in the target source to form a representative thematic text collection (text dataset) of large dimensions, being a special case of obtaining and structuring primary data, remains one of the most demanded applied tasks of natural language processing. The article presents the experience of developing a system of linguistic markers that allows automated extraction of texts related to the topic of vaccination against COVID-19 on the material of the VKontakte social network. A combination of linguistic methods with methods for collecting and processing text data allows forming the final dataset. The test list of markers forms is based on background knowledge, work with dictionaries and special linguistic services. The task was to create a list of words united by a common conceptual feature, to predict the joint occurrence of words in texts about vaccination against COVID-19, or to find specific words that mark this topic: occasionalisms, designations of specific realities. The content of the VKontakte thematic communities uploaded using the test list of markers became the source of automated and expert extraction of the main array of markers (354 units). The procedure for automated filtering of an intermediate text sample (12.8 million texts) is in detail. The technique of formation of stop-words is given. For the period from 01.01.2020 to 03.01.2023, 4.5 million relevant messages were retrieved; the validity of the markers was confirmed by an insignificant amount of noise on the scale of big data. The general principles of preparing linguistic markers for automated unloading of large text data are systematized; the strengths and weaknesses of this tool are noted; recommendations for the formation of a list of linguistic markers are suggested.

Keywords
Linguistic marker, big data, automated data collection, data upload, text collection, full-text search, social networks, VK, open API, vaccination, COVID-19.

Введение

В исследованиях социальных явлений и процессов на материале больших данных, которые становятся все более заметными как в зарубежном, так и в отечественном научном поле, первым и, как правило, обязательным техническим этапом является этап сбора данных. В случае текстовых данных — а значительная доля данных в сети Интернет имеет текстовый формат — возникает необходимость подготовить репрезентативную текстовую коллекцию, на материале которой будет осуществляться анализ данных. Критерием отбора текстов часто становится тематический: на материале больших текстовых коллекций, сформированных по тематическому принципу, изучаются в актуальном временном срезе, например, такие важные социальные явления, как субъективное качество жизни населения [Shchekotin et al. 2021], факторы, влияющие на склонность населения к профилактическим мероприятиям для защиты здоровья [Huang et al. 2020], отношение к планированию семьи [Deng et al. 2021], экстремистская деятельность [Ahmad et al. 2019] и др.

В практиках в области обработки естественного языка известны два основных способа автоматизированного формирования тематической текстовой коллекции: с помощью лингвистических маркеров и с помощью обучающей выборки. В первом случае текст попадает в выборку, если он содержит заданный маркер. Во втором — готовится экспертно размеченная вручную коллекция текстов, на основании которой алгоритмы ищут похожие тексты в определенном источнике данных: в основе операций сравнения, классификации, кластеризации текстов оказывается не детекция формального маркера, а семантическое сходство текстов.

Главное преимущество поиска с помощью обучающей выборки заключается в меньшей степени субъективизма при получении результата (но не абсолютной, так как тексты, отобранные в обучающую выборку, также не могут охватить все семантическое пространство заданной тематики). Однако поиск текстов с помощью обучающей выборки трудоемкий, дорогой, требующий высокого качества разметки, заточенный под конкретную тематику (модель нельзя использовать для других тематик), а результаты также требуют серьезной валидации, поэтому метод достаточно редко используется именно на стадии сбора первичных данных. Чаще всего данное решение становится результатом специального исследования, а не вспомогательным этапом.

Лингвистические маркеры, таким образом, сегодня остаются основным инструментом автоматизированного отбора текстов определенной тематики из всей совокупности текстов интернет-источника. При этом данный инструмент требует ряда сопроводительных процедур и по факту обычно используется как одна из фаз выгрузки данных (см., например, [Карпова и др. 2020]).

Маркером называется знак, наличие которого свидетельствует об определенных свойствах, признаках, функциях более крупной единицы, частью которой он является. В лингвистике маркерами могут выступать морфемы, лексемы, формализованные аспекты значения (например, особенности коннотации), на синтаксическом уровне — маркеры модусных и диктумных смыслов и др.

Для автоматизированного поиска и отбора текстов по заданной теме наиболее востребованными оказываются лексические маркеры: слова и их комбинации.

В научном дискурсе встречаются термины «лингвистические маркеры» [Cohen et al. 2014; Мишланов и др. 2020; Карпова и др. 2020; Erseghe et al. 2022], «психолингвистические маркеры» [Сбоев и др. 2013], «вербальные маркеры» [Колмогорова и др. 2019][3], «ключевые слова» [Карпова и др. 2020], «языковые маркеры» [Колмогорова и др. 2016], «лингвистические предикторы» [Liu et al. 2022]. Близкими оказываются и термины из области информационного поиска: «поисковые запросы», «маркерные запросы», «семантические запросы», «семантические маркеры». Размытость и неконвенциональность термина уже констатировалась исследователями ранее [Горностаева 2018].

При этом можно наблюдать, что термин «вербальные маркеры» чаще используется в контексте отбора текстов не по тематическому, а по лингвопрагматическому критерию (манипулятивные тексты, разного рода эмотивные характеристики и др.). Термины «поисковые запросы», «маркерные запросы», «семантические запросы», «семантические маркеры» обычно связаны с запросами пользователей в поисковых системах. Термин «психолингвистические маркеры» является более узким и связан с психологическим состоянием коммуниканта в момент продуцирования текста. Термин «ключевые слова» (keywords) в англоязычной научной литературе, как правило, относится к операции извлечения ключевых слов (keyword extraction), а не к поиску по словам (см., например, [Huh 2018]).

Под лингвистическим маркером будем понимать в данной работе слово или комбинацию слов, наличие которых в тексте будет свидетельствовать о принадлежности данного текста к текстам определенной тематики (в нашем случае к дискурсу вакцинации от COVID-19).

В статье решены следующие задачи: систематизированы принципы разработки списка лингвистических маркеров; рассмотрен пример формирования системы лингвомаркеров по теме вакцинации от COVID-19 и автоматизированной выгрузки тематических текстов из социальной сети «ВКонтакте» за период с 1 января 2020 г. по 1 марта 2023 г.; предложены рекомендации по формированию лингвистических маркеров.

Актуальность работы обусловлена значимостью больших данных и социальных сетей как источника современных эмпирических изысканий в социальных науках; востребованностью инструментария лингвистических маркеров и слабой освещенностью данной темы в качестве самостоятельного исследования; практическим характером работы, уточнением методологии автоматизированного сбора неструктурированных данных; актуальной проблематикой вакцинации от COVID-19.

[3] См. также Горностаева Ю.А. Вербальные маркеры манипуляции в англоязычном поляризованном политическом дискурсе: опыт параметризации и автоматической обработки: Автореферат дис. … канд. филол. наук. Красноярск, 2018.

Материал и методы

Источником данных выбрана социальная сеть «ВКонтакте» по причине ее популярности среди российских интернет-пользователей (97 млн посетителей ежемесячно), а также открытого API, отличающего ее от ряда других социальных сетей и позволяющего свободно извлекать данные в исходном виде. Извлекались только открытые данные, все данные обезличены. Контент извлекался за период с 01.01.2020 по 01.03.2023, то есть с момента распространения пандемии COVID-19 по настоящее время.

Использовались следующие общенаучные методы исследования: анализ, синтез, обобщение, классификация, описание, эксперимент. Использовались также методы лингвистического анализа: работа с лексикографическими источниками, лексический и словообразовательный анализ слова, оценка тематической значимости лексемы, элементы дистрибутивного анализа. Кроме того, применялись следующие методы сбора и обработки данных: выгрузка с помощью API, фильтрация нерелевантных данных, статистический, валидация.

Для автоматизированной обработки данных использовалась платформа текстовой аналитики PolyAnalyst (разработчик — российская компания Megaputer Intelligence), доступная через Центр коллективного пользования Национального исследовательского Томского государственного университета [Петров, Саркисова 2021].

Общие принципы подготовки лингвистических маркеров

Задача — сформировать список маркеров, которые позволят не потерять значимые тексты, но при этом не собрать лишние.

Сама специфика лексической системы языка (неограниченность количества слов, зыбкость границ между лексико-тематическими группами, темпоральная подвижность значений, проницаемость) делает задачу автоматического отбора текстов по словам достижимой только в весьма относительной степени: лексика — самая открытая система языка и в наибольшей степени зависимая от внеязыковых факторов. Поэтому реальная задача при отборе тематических текстов состоит лишь в получении максимально большой, репрезентативной выборки.

Обозначим основные ограничения лингвомаркеров как инструмента сбора больших текстовых данных:

1. маркеры формируются исследователем, поэтому изначально субъективны, не абсолютны и не претендуют на полную идентификацию тематического контента;
2. неизбежно частичное попадание в выборку нерелевантного контента. Маркеры могут характеризоваться полисемантичностью, иметь омонимы, входить в состав разного рода онимов (названия фильмов, товаров и т.д., что является часто обсуждаемым в соцсетях контентом).
3. при использовании выборки релевантных текстов как источника извлечения новых лингвомаркеров необходимо иметь в виду, что каждое слово, помимо актуальных сем (реализуемых в конкретном контексте), обладает трудно прогнозируемым набором потенциальных сем (которые возможно учесть, только просматривая другие контексты).

Несмотря на названные ограничения, принципиальная возможность разделить слова по семантическим полям, с одной стороны, и достижения дистрибутивной семантики, с другой стороны, позволяют успешно решать задачи автоматизированного формирования тематических текстовых коллекций большой размерности.

Выработаны следующие меры по предупреждению ошибок в составлении списка лингвомаркеров, рекомендации и процедурные этапы.

1. Маркеры подвергаются тщательному лингвистическому анализу на предмет полисемии, омонимии, особенностей словообразования и словоизменения, функциональности (употребляемость, частотность, темпоральность, стилистическая окраска и др.), парадигматических связей, включенности в устойчивые словосочетания и прецедентные высказывания, ограничений в сочетаемости. Рекомендуется избегать маркеров, представляющих частотные в языке слова, которые могут встретиться в текстах любой тематики. При необходимости включения подобного слова в список маркеров рекомендуется приводить его в составе словосочетаний или в паре с другим словом (словами), которые должны присутствовать в тексте одновременно. Количество таких наборов слов технических ограничений не имеет. Грамматическая связь слов и непосредственное соседство в тексте также не являются обязательными условиями.

Другие маркеры, напротив, будучи слишком узкими (имеют специализированные значения, стилистические ограничения, окказиональное происхождение, обозначают уникальные реалии (например, названия вакцин) и т.д.), часто удобны и рекомендуются к включению для точной идентификации целевого контента, но должны быть дополнены другими маркерами во избежание потери существенной части контента, не содержащего их.

Таким образом, учет средней частоты употребления слова в языке является одним из ключевых факторов (соотносится с метрикой тематической значимости слова в теории обработки естественного языка). Анализ значимости слов в текстовой коллекции может представлять и исследовательский интерес с точки зрения «оценки близости и глубины пересечения индивидуальных лексиконов с обобщенным лексиконом сообщества» [Карпова 2020, 170].

2. Необходимо тщательно прорабатывать список стоп-слов, в том числе словосочетаний и комбинаций, одновременно встречающихся в тексте слов. Стоп-слово — это маркер, при наличии которого текст в искомую выборку не попадет.

3. Источниками маркеров в первой итерации, помимо фоновых знаний исследователя по искомой тематике, могут служить разного рода словари, научная литература, запросы к экспертам, предварительное выборочное ознакомление с целевыми источниками (если источником выступают социальные сети, полезно предугадывать маркеры, соответствующие жанрово-стилевым параметрам свойственной им коммуникации), а также интернет-сервисы.

С одной стороны, формируются тематические группы слов на основе лексем, объединенных единым концептуальным признаком. С другой стороны, полезно спрогнозировать совместное употребление слов в общих контекстах. Синтагматическое структурное значение слова характеризует совокупность всех окружений, в которых употребляется данная лексическая единица, то есть ее дистрибуция. Оно опирается на сочетаемость смыслов (наличие в понятийных значениях слова общих сем). Однако в данном случае интерес представляет не только непосредственная (в пределах двух слов) семантическая валентность слова, а множество всех окружений (смысловых контекстов), в которых встречается некоторая лексическая единица.

Для подбора слов, которые часто могут встречаться в одних контекстах, то есть маркировать одну и ту же тематику, существуют специальные сервисы. Для работы с русскоязычными текстами, в частности, могут быть рекомендованы сервисы «RusVectōrēs: семантические модели для русского языка» (см. о нем также [Концевой 2022]) и «Картаслов.ру». Сервис RusVectōrēs основан на закономерностях дистрибутивной семантики и позволяет устанавливать

семантически близкие слова при помощи дистрибутивных моделей: семантически связанные слова вычислены на основе совместной встречаемости в схожих контекстах. Сервис «Картаслов.ру» позволяет увидеть примеры парадигматических (синонимы) и синтагматических (сочетаемость, примеры употреблений в контексте) связей слова, изучить сеть словесных ассоциаций для построения номинативного поля концепта.

Могут быть также полезны сервисы, показывающие частоту поисковых запросов по некоторой теме. Отслеживая поисковый спрос, сервисы одновременно обнаруживают ключевые слова, обозначающие актуальные для пользователей предметы и явления. Сервис «ЯндексВордстат» при вводе пользователем слова или словосочетания показывает статистику запросов в поисковой системе «Яндекс» данного слова и похожих запросов. Сервис особенно удобен для формирования списка стоп-слов, позволяя увидеть однозначно нерелевантные словосочетания с искомым словом (например, одноименные названия фильмов или видеоигр). Сервис GoogleTrends показывает графики, визуализирующие то, как часто пользователи поисковой системы Google ищут данное слово по отношению к общему объему поисковых запросов в динамике. Сервис позволяет сравнивать частоту конкретных поисковых запросов (например, «вакцинация от ковида» и «вакцинация от гриппа») (см., например, [Kessel et al. 2023]).

4. Первый список лингвомаркеров носит, как правило, тестовый характер. На основе тестовых лексических маркеров извлекается контент, который, с одной стороны, частично просматривается вручную — список маркеров и стоп-слов корректируется; с другой стороны, из контента автоматизированно извлекаются ключевые слова (наиболее частотные и значимые одновременно) — список лингвомаркеров обновляется и за счет них. На основе обновленных списков лингвомаркеров и стоп-слов контент выгружается повторно. Процедура повторяется до тех пор, пока есть прирост качества.

5. Извлечение контента с помощью API позволяет отследить, какой маркер идентифицировал конкретное сообщение. На основе случайной выборки маркеры могут быть подвергнуты ручной валидации: исследователь проверяет, насколько релевантный контент попал в выборку по каждому маркеру в отдельности. Имеет место статистическая экспертиза валидности выявленных маркеров.

Опыт формирования системы лингвистических маркеров для выгрузки контента из социальной сети «ВКонтакте» по тематике вакцинации от COVID-19 за период с 01.01.2020 по 01.03.2023

Полная осуществленная процедура разработки системы лингвистических маркеров, выгрузки, фильтрации и валидации целевых текстовых данных отражена на Рисунке 1.

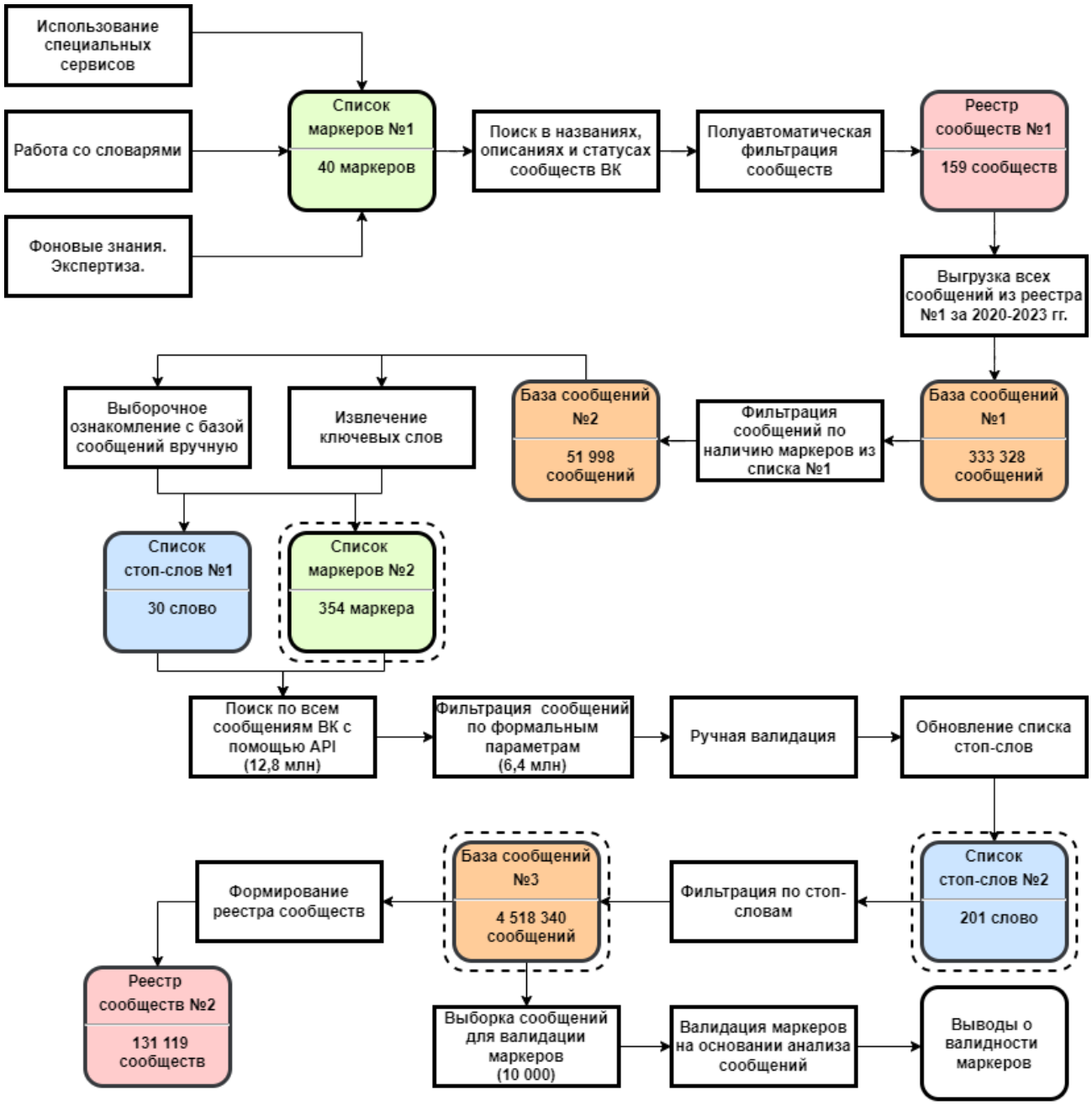

Рисунок 1. Схема разработки системы лингвомаркеров, извлечения и фильтрации тематических текстов из социальной сети «ВКонтакте»[4]

Описание процедуры формирования списка лингвомаркеров. На первоначальном этапе стояла задача погрузиться в предметное поле исследования и сформировать первичный (тестовый) список лингвистических маркеров по тематике вакцинации от COVID-19 для идентификации и выгрузки релевантных текстовых данных из социальной сети «ВКонтакте».

Для составления первичного списка лингвомаркеров была проведена выборочная экспертная оценка тематического контента в социальных медиа, работа с аналитическими сервисами «ЯндексВордстат» и GoogleTrends, с помощью которых были выявлены наиболее часто используемые слова и словосочетания по тематике вакцинации в поисковых запросах, а также работа со словарями.

[4] Составлено авторами на основе проведенного исследования.

В частности, использовался «Словарь русского языка коронавирусной эпохи»[5] — собрание пандемийных неологизмов 2020–2021 гг. (ок. 3500 единиц), выявленных составителями на основе мониторинга русскоязычных СМИ и Интернета. Словарь является историческим, привязан к короткому периоду (2020–2021 гг.). Он оптимален как дополнительный источник, содержащий некоторые точные узкие маркеры (барановирус, ковидобесие и др.). Необходимо, однако, как и при составлении списка маркеров в целом, иметь в виду обилие в источнике слов-«ярлыков» (например, антиваксер, мракобес, вакцинатор, ковидятел), маркирующих оценку оппонентов, а не тексты самих сторонников / противников вакцинации, и предупредить уклон в сторону преобладания подобных «легких» маркеров.

Сервис «ЯндексВордстат» позволил обнаружить комбинации, точно маркирующие дискурс вакцинации от COVID-19 (вакцинация & госуслуги; вакцинация & ковид & купить и др.), а также некоторые стоп-слова (вакцинация щенков, вакцинация & АКДС и др.).

Был составлен тестовый список лингвомаркеров, в который вошли 40 слов и комбинаций слов. По отобранным лингвомаркерам был осуществлен полнотекстовый поиск не по контенту, а сначала по названиям, описаниям и статусам всех сообществ социальной сети «ВКонтакте».

Одновременно сообщества прошли процедуру автоматической фильтрации для предварительного исключения сообществ, нерелевантных задачам исследования. Критерии фильтрации: 1) сообщество имеет 1000 и более подписчиков; 2) сообщество не является «событием/мероприятием» (event). В результате в первичный реестр сообществ вошли 159 сообществ в сети «ВКонтакте» — коммуникативных площадок по обсуждению вакцинации. Примеры извлеченных сообществ: «Ваша мама в секте», «Прививка от мракобесия 18+», «Все о Чипизации и Биометризации», «ПВО — поствакцинальные осложнения» и др.

Далее с помощью открытого API «ВКонтакте» были выгружены все сообщения из первичного реестра сообществ за 2020–2023 гг. (333 328 постов). Полученные сообщения проверялись на наличие тестовых лингвомаркеров. Если в сообщении встречался хотя бы один из лингвомаркеров, оно включалось в базу тематических сообщений. Таким образом, был сформирован датасет из 51 998 сообщений.

Для расширения списка лингвомаркеров и фильтрации нерелевантных постов из этого датасета автоматически с помощью узла «Извлечение ключевых слов» аналитической системы PolyAnalyst извлекались ключевые слова (слова с наибольшим весом по отношению к остальным словам в текстах). Посредством ручной валидации часть извлеченных слов и словосочетаний добавлялась в список лингвистических маркеров (например, при необходимости в комбинациях с другими словами: Минздрав, цифровой пропуск, Роспотребнадзор, штамм, псевдоученый и др.).

Параллельно велось выборочное экспертное ознакомление с выгруженными текстами, на основании их анализа извлекались новые лингвомаркеры, а также был составлен первый список стоп-слов (30 единиц). Изучение контента «вручную» спровоцировало такие новые маркеры (при необходимости — в комбинациях с другими словами), как биооружие, биологический материал, коллективный иммунитет, ИВЛ, тайный мировой, фармделец, невежество, вакханалия и др. Таким образом, был разработан основной список лингвистических маркеров (354 единицы). Ниже приведена его характеристика.

Характеристика основного списка лингвистических маркеров. Слово вакцина относится к терминологической (медицинской), специальной лексике, поэтому отличается конкретной семантикой. У него нет общеузуальных производных значений, оно достаточно

[5] Словарь русского языка коронавирусной эпохи / Сост. Х. Вальтер, Е.С. Громенко, А.Ю. Кожевников и др. Санкт-Петербург: Институт лингвистических исследований РАН, 2021.

слабо подвержено метафоризации. Можно предположить, что слово будет маркировать именно тексты, связанные с тематикой вакцинации. Это же касается производных от слова вакцина слов (*вакцинировать, вакцинироваться, вакцинация, вакцинатор, антивакцинатор, вакцинный* и т.д.). При словоизменении и словообразовании фузии на стыке морфем, а также внутренней флексии не наблюдается, поэтому для всей группы слов данного словообразовательного гнезда в качестве инварианта возможно выбрать основу «вакцин». (Особняком стоят возникшие в результате калькирования слова *ваксер, антиваксер* и их русские разговорные дубликаты *вакер, провакер, антивакер*. Эти слова и их производные имеют другой корень, поэтому должны быть включены в список маркеров отдельно.)

Среди синонимов слова *вакцина* широкоупотребительным является слово *прививка*, оно должно быть учтено при поиске как равнозначное. Однако глагол прививать (в отличие от *вакцинировать*) является многозначным и частотным, поэтому не может служить удачным маркером. В данном случае в качестве основного маркера приходится выбрать основу «прививк», при этом не соответствующие ей слова данного словообразовательного гнезда (*прививаться, антипрививочник* и др.) включать позже в качестве отдельных маркеров, при необходимости — в конструкциях (*прививать* от и др.), одновременно вводя в список стоп-слов словосочетания типа *прививать уважение, прививать интерес* и др.

Задача заключалась в том, чтобы выгрузить данные, связанные с вакцинацией именно от COVID-19. Поэтому, с одной стороны, основы «вакцин» и «прививк» требуют уточнений при формировании поискового запроса. С другой стороны, исключить попадание в выборку нерелевантных данных помогают стоп-слова. Существенным внеязыковым фактором выступил и период, за который было необходимо собрать данные, — период пандемии: обсуждение вакцинации от COVID-19 в это время является доминирующим по отношению к другим болезням. В качестве уточнений добавлены маркеры, полученные на основе разного рода источников (словари, сервисы, ключевые слова, предварительное знакомство с контентом и др.). 30 сформированных стоп-слов составили в первую очередь слова, отсылающие к вакцинации животных, и названия ряда других болезней (не коронавирусная инфекция).

Итоговые маркеры данной группы имеют вид поискового запроса: «*вакцин или прививк & концлагерь*», «*вакцин или прививк & всеобщий*», «*вакцин или прививк & принуждать*», «*вакцин или прививк & бизнес*» и т.д. Использовано 308 маркеров, построенных по данной схеме.

Во избежание потери контента, релевантного теме, но не включающего элементы «вакцин» и «прививк», была подготовлена вторая группа маркеров. Это слова, которые будут маркировать искомый контент в любых контекстах (изолированные маркеры). В данном случае это названия вакцин от COVID-19 (*Sputnik V, гам-ковид-вак* и др.), некоторые окказионализмы (*барановирус, фармфашизм, вакцинопроповедник, сПутник* и др.). Использовано 22 маркера.

Третью группу маркеров составили слова, интересные авторам с точки зрения исследовательских задач, но не способные выполнять функцию фильтра ввиду многозначности, частотности и широкой употребляемости в нерелевантных контекстах. Данные слова были включены в конструкции не только с основами «вакцин» или «прививк», но и одновременно с основами «ковид», или «корон», или «covid». Таким образом, итоговый вид маркера следующий: «*прививк или вакцин & ковид или корон или covid & американский*», «*прививк или вакцин & ковид или корон или covid & ребенок*», «*прививк или вакцин & ковид или корон или covid & антинаучный*» и т.д. Подготовлено 24 маркера. Общий список маркеров составил 354 единицы.

Извлечение текстовых данных и их фильтрация. С помощью методов API «ВКонтакте» был произведен поиск всех сообщений, опубликованных в период 01.01.2020–01.03.2023, включающих хотя бы один из 354 маркеров и не содержащих при этом ни одного из 30 стоп-слов. Таким образом, методом полнотекстового поиска было обнаружено и извлечено 12,8 млн сообщений. Их распределение по годам отражает актуальность темы вакцинации от COVID-19 в России, пик которой пришелся на 2021 год, когда уже имелась выпущенная в оборот вакцина и в стране стартовала массовая вакцинация: 2020 — 2,3 млн, 2021 — 7,8 млн, 2022 — 2,5 млн, 2023 — 202 тыс. сообщений. Дальнейшая обработка текстов осуществлялась с помощью аналитической платформы PolyAnalyst (Рисунок 2).

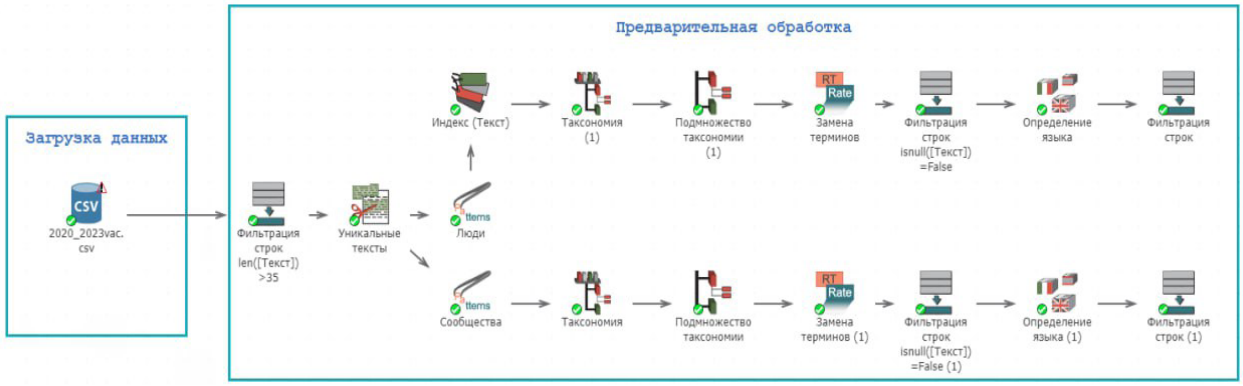

Рисунок 2. Этапы автоматической фильтрации выгруженных текстовых данных на платформе PolyAnalyst[6]

Стояла задача удалить сообщения, не отвечающие исследовательским целям: посты без текста, повторяющиеся посты, однословные и несодержательные по причине лаконичности, тексты на иностранных языках (попавшие в выборку из-за маркеров на латинице: Gam-COVID-Vac, Sputnik V и т.п.), посты, состоящие только из хештегов или ссылок.

После удаления пустых значений и абсолютных дублей осталось 6,4 млн сообщений: сокращение выборки сразу наполовину отражает факт огромной доли «несамостоятельного» контента, свойственного социальным сетям в целом, в том числе вбросов и репостов.

Далее были удалены сообщения меньше 35 символов — на основе построения распределения предварительно было установлено, что такие посты малоинформативны для исследовательских задач; после их фильтрации осталось 6,3 млн текстов.

Следующим этапом стало удаление текстов, схожих более чем на 80% (сохранялся самый длинный): это скопированные тексты, в которые пользователь добавлял незначительное изменение — например, ссылку на искомый текст. После данного этапа осталось 6,1 млн текстов.

Полученные тексты были разделены на тексты, опубликованные от лица сообщества, и тексты, опубликованные непосредственно пользователями. Разделение связано с дальнейшими целями исследования, а также с целью сокращения вычислительной нагрузки каждой дальнейшей операции. Выявлено, что 1 692 000 текстов было опубликовано от лица сообществ, 4 430 000 — от лица пользователей.

[6] Составлено авторами на основе проведенного исследования.

Далее стояла задача удаления тематически нерелевантного контента. Для этого был обновлен список стоп-слов следующим образом. Была создана случайная выборка из 10 000 сообщений: по 2 000 сообщений за каждый год, за исключением наиболее активного 2021 года, по которому было отобрано 4000 сообщений. Выборка была отсмотрена вручную. Экспертно выделялись слова и словосочетания, генерирующие шум. Например, обнаружено, что изначально выбранное в качестве лингвистического маркера слово *ПЦР* сгенерировало большой массив рекламного контента о поездках и путешествиях. Изначальное включение стоп-слов *собака, пес, щенок* оказалось явно недостаточным: выборка показала, что необходимо также перечислять отдельные породы. Обилие постов о вакцинации животных в сети, как обнаружилось, связано прежде всего с многочисленными объявлениями о продаже породистых животных, включающими указание на уже сделанные прививки. Объявления о купле-продаже всегда составляют большой процент контента социальной сети, их фильтрация должна быть тщательно продумана применительно к каждой конкретной теме. Термин полиомиелит, изначально введенный в список стоп-слов, как оказалось, имеет в текстах пользователей множество вариантов написания, что в существенной степени нейтрализовало данное стоп-слово. Выявлялись также случайные генерирующие шум слова. Список стоп-слов пополнился на текущем этапе на 171 слово. На основе обновленного списка стоп-слов (201 единицы) посты подвергались повторной фильтрации, после чего осталось 4 557 315 сообщений (из них 1 081 125 постов от лица сообществ, 3 476 190 постов от лица пользователей).

Последним этапом стало удаление ссылок, хештегов, текстов на иностранных языках. С помощью регулярных выражений были составлены правила для поиска ссылок и хештегов, после чего найденные в текстах ссылка или хештег заменялись на пустые значения. Далее повторялась процедура удаления пустых значений и постов меньше 35 символов (так как после удаления ссылок длина текстов изменилась). Заключительной операцией фильтрации стало определение языка текста и фильтрация всех, кроме русского.

Таким образом, итоговая текстовая коллекция составила 4 518 340 сообщений (в том числе 1 066 759 текстов от лица сообществ, 3 451 581 текст от лица пользователей).

Данные сообщения оказались размещены на страницах 131 119 сообществ и на страницах 1 145 360 пользователей «ВКонтакте».

Валидация маркеров. Для валидации маркеров из итоговой коллекции текстов (4 518 340 единиц) было рандомно отобрано 10 000 сообщений (по 2000 за каждый год, за 2021 год — 4000). В данной выборке был осуществлен поиск каждого из 354 маркеров, тексты были сгруппированы по критерию включенности каждого маркера (если текст включал два и более маркеров, он отображался в каждой группе). Просмотр текстов по каждому маркеру подтвердил несущественное количество шума.

Полученная статистика частотности маркеров в социальной сети показала, что наиболее популярными (Рисунок 3) оказались маркеры, отсылающие к вакцинации от коронавируса в общем; большое количество контента посвящено вакцине российского производства «Спутник V»; много текстов содержит маркеры о вакцинации в сочетании со словами лексико-тематической группы «Здоровье» (*заболеть, умереть, здоровье* и др.). Некоторые популярные маркеры диагностируют ведущие нарративы: качество испытаний / исследований вакцины, доверие к власти, право на выбор, поствакцинальные осложнения и последствия, наличие антител и необходимость вакцинации, проблемы российского здравоохранения, эффективность/ неэффективность вакцинации и др.

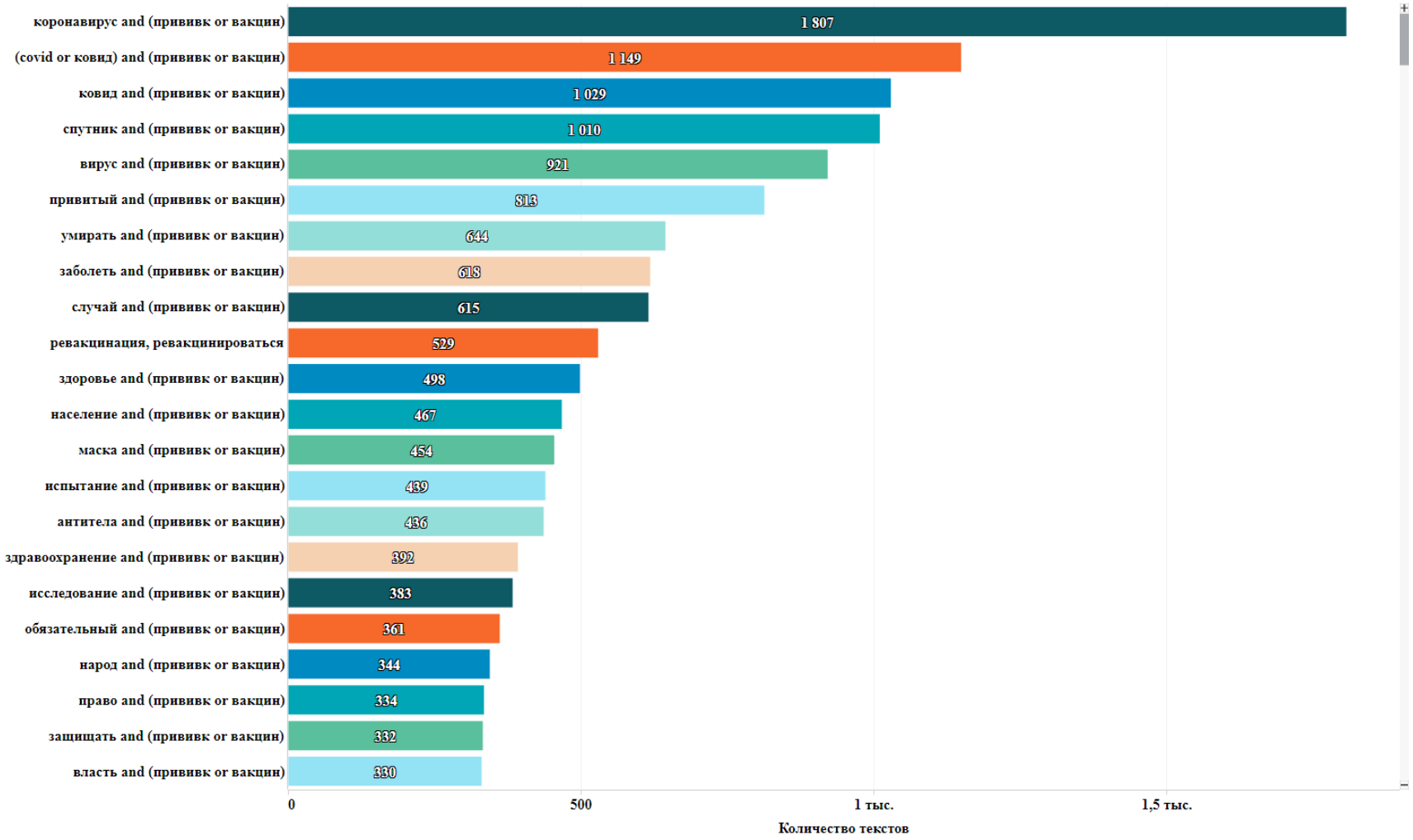

Рисунок 3. Маркеры, сгенерировавшие наибольшее количество сообщений в итоговой текстовой коллекции[7]

[7] Составлено авторами на основе проведенного исследования.

Заключение

Представленный пример использования лингвистических маркеров в качестве инструмента автоматизированного сбора неструктурированных тематических данных отражает основные стадии, возможности и ограничения реализации данного инструмента для решения подобного рода задач.

Сформированная текстовая коллекция (4,5 млн сообщений) будет использована в качестве эмпирических данных для всестороннего аналитического исследования общественного отношения и поляризации мнений о вакцинации от COVID-19 в российском цифровом обществе.

Список литературы:

Горностаева Ю.А. Опыт выявления вербальных маркеров психологических и когнитивных процессов в лингвистике: к истории вопроса // Филологические науки. Вопросы теории и практики. 2018. № 8(86). Ч. 1. С. 91–94. DOI: 10.30853/filnauki.2018-8-1.21

Карпова А.Ю., Савельев А.О., Вильнин А.Д., Чайковский Д.В. Изучение процесса онлайн-радикализации молодежи в социальных медиа (междисциплинарный подход) // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 3. С.159–181. DOI: 10.14515/monitoring.2020.3.1585

Колмогорова А.В., Талдыкина Ю.А., Калинин А.А. Языковые маркеры манипуляции в поляризованном политическом дискурсе: опыт параметризации // Политическая лингвистика. 2016. № 4(58). С. 194–199.

Колмогорова А.В., Калинин А.А., Маликова А.В. Типология и комбинаторика вербальных маркеров различных эмоциональных тональностей в интернет-текстах на русском языке // Вестник Томского государственного университета. 2019. № 448. С. 48–58. DOI: 10.17223/15617793/448/6

Концевой М.Р. Онлайновые семантические вычисления на платформе RusVectōrēs в преподавании компьютерной лингвистики // Дистанционное обучение — образовательная среда XXI века: материалы XII Международной научно-методической конференции, Минск, 26 мая 2022 г. Минск: БГУИР, 2022. С. 75.

Мишланов В.А., Каджая Л.А., Кузнецова Ю.М. Лингвистические маркеры эмоционального состояния субъекта речи (к проблеме автоматического мониторинга текстов сетевой коммуникации) // Медиалингвистика. 2020. Т. 7. № 4. С. 428–444. DOI: 10.21638/spbu22.2020.405

Петров Е.Ю., Саркисова А.Ю. Ресурс аналитической платформы PolyAnalyst в социогуманитарных научных исследованиях // Открытые данные — 2021: материалы форума / под ред. А.Ю. Саркисовой. Томск: Издательство Томского государственного университета, 2021. С. 94–104.

Сбоев А.Г., Гудовских Д.В., Молошников И.А., Кукин К.А., Рыбка Р.Б., Иванов И.И., Власов Д.С. Автоматическое выделение психолингвистических характеристик текстов в рамках концепции Big Data // Современные информационные технологии и IT-образование. 2013. № 9. С. 433–438.

Ahmad S., Asghar M.Z., Alotaibi F.M., Awan I. Detection and Classification of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques // Human-centric Computing and Information Sciences. 2019. Vol. 9. DOI: 10.1186/s13673-019-0185-6

Cohen K., Johansson F., Kaati L., Clausen Mork J.C. Detecting Linguistic Markers for Radical Violence in Social Media // Terrorism and Political Violence. 2014. Vol. 26. Is. 1. P. 246–256. DOI: 10.1080/09546553.2014.849948

Deng W., Hsu J.-H., Löfgren K., Cho W. Who Is Leading China's Family Planning Policy Discourse in Weibo? A Social Media Text Mining Analysis // Policy & Internet. 2021. Vol. 13. Is. 4. P. 485–501. DOI: 10.1002/poi3.264

Erseghe T., Badia L., Dzanko L., Suitner C. PLMP: A Method to Map the Linguistic Markers of the Social Discourse onto Its Semantic Network // 2022 IEEE / ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). November 10–13, 2022, Istanbul, Turkey. Istanbul: Institute of Electrical and Electronics Engineers, 2022. P. 247–251. DOI: 10.1109/ASONAM55673.2022.10068643

Huang F., Ding H., Liu Z., Wu P., Zhu M., Li A., Zhu T. How Fear and Collectivism Influence Public's Preventive Intention towards COVID-19 Infection: A Study Based on Big Data from the Social Media // BMC Public Health. 2020. Vol. 20. DOI: 10.1186/s12889-020-09674-6

Huh J-H. Big Data Analysis for Personalized Health Activities: Machine Learning Processing for Automatic Keyword Extraction Approach // Symmetry. 2018. Vol. 10. Is. 4. DOI: 10.3390/sym10040093

Kessel R. van, Kyriopoulos I., Wong B.L.H., Mossialos E. The Effect of the COVID-19 Pandemic on Digital Health–Seeking Behavior: Big Data Interrupted Time-Series Analysis of Google Trends // Journal of Medical Internet Research. 2023. Vol. 25. DOI: 10.2196/42401

Liu T., Giorgi S., Yadeta K., Schwarts H.A., Ungar L.H., Curtis B. Linguistic Predictors from Facebook Postings of Substance Use Disorder Treatment Retention versus Discontinuation // The American Journal of Drug and Alcohol Abuse Encompassing. 2022. Vol. 48. Is. 5. P. 573–585. DOI: 10.1080/00952990.2022.2091450

Shchekotin E.V., Goiko V.L., Myagkov M.G., Dunaeva D.O. Assessment of Quality of Life in Regions of Russia Based on Social Media Data // Journal of Eurasian Studies. 2021. Vol. 12. № 2. DOI: 10.1177/18793665211034185

References:

Ahmad S., Asghar M.Z., Alotaibi F.M., Awan I.(2019) Detection and Classification of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques. *Human-centric Computing and Information Sciences.* Vol. 9. DOI: 10.1186/s13673-019-0185-6

Cohen K., Johansson F., Kaati L., Clausen Mork J.C. (2014) Detecting Linguistic Markers for Radical Violence in Social Media. *Terrorism and Political Violence.* Vol. 26. Is. 1. P. 246–256. DOI: 10.1080/09546553.2014.849948

Deng W., Hsu J.-H., Löfgren K., Cho W.(2021) Who Is Leading China's Family Planning Policy Discourse in Weibo? A Social Media Text Mining Analysis. *Policy & Internet.* Vol. 13. Is. 4. P. 485–501. DOI: 10.1002/poi3.264

Erseghe T., Badia L., Dzanko L., Suitner C. (2022) PLMP: A Method to map the linguistic markers of the social discourse onto its semantic network. 2022 IEEE / ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). November 10–13, 2022, Istanbul, Turkey. 2022. Istanbul: Institute of Electrical and Electronics Engineers. P. 247–251. DOI: 10.1109/ASONAM55673.2022.10068643

Huang F., Ding H., Liu Z., Wu P., Zhu M., Li A., Zhu T. (2020) How Fear and Collectivism Influence Public's Preventive Intention towards COVID-19 Infection: A Study Based on Big Data from the Social Media. *BMC Public Health.* Vol. 20. DOI: 10.1186/s12889-020-09674-6

Huh J-H. (2018) Big Data Analysis for Personalized Health Activities: Machine Learning Processing for Automatic Keyword Extraction Approach. *Symmetry.* Vol. 10. Is. 4. DOI: 10.3390/sym10040093

Gornostaeva Yu.A. (2018) Attempt of Identifying Verbal Markers of Psychological and Cognitive Processes in Linguistics: On the Issue History. *Filologicheskie nauki. Voprosy teorii i praktiki.* No. 8(86). Part 1. P. 91–94. DOI: 10.30853/filnauki.2018-8-1.21

Karpova A.Yu., Savelev A.O., Vilnin A.D., Chaykovskiy D.V. (2020) Studying Online Radicalization of Youth through Social Media (Interdisciplinary Approach). *Monitoring obshchestvennogo mneniya: ekonomicheskiye i sotsial'nyye peremeny.* No. 3. P. 159–181. DOI: 10.14515/monitoring.2020.3.1585

Kessel R. van, Kyriopoulos I., Wong B.L.H., Mossialos E. (2023) The Effect of the COVID-19 Pandemic on Digital Health–Seeking Behavior: Big Data Interrupted Time-Series Analysis of Google Trends. *Journal of Medical Internet Research*. Vol. 25. DOI: 10.2196/42401

Kolmogorova A.V., Taldykina Yu.A., Kalinin A.A. (2016) Linguistic Markers of Manipulation in Polarized Discourse: Parametric Study. *Politicheskaya lingvistika.* No. 4(58). P. 194–199.

Kolmogorova A.V., Kalinin A.A., Malikova A.V. (2019) The Types and Combinatorics of Verbal Markers of Different Emotional Tonalities in Russian-Language Internet Texts. *Vestnik Tomskogo gosudarstvennogo universiteta.* No. 448. P. 48–58. DOI: 10.17223/15617793/448/6

Kontsevoy M.P. (2022) Onlaynovyye semanticheskiye vychisleniya na platforme RusVectōrēs v prepodavanii komp'yuternoy lingvistiki [Online semantic calculations on the RusVectōrēs platform in teaching computational linguistics]. *Distantsionnoye obucheniye — obrazovatel'naya sreda XXI veka: materialy XII Mezhdunarodnoy nauchno-metodicheskoy konferentsii.* Minsk, May 26, 2022. Minsk: BGUIR. P. 75.

Liu T., Giorgi S., Yadeta K., Schwarts H.A., Ungar L.H., Curtis B. (2022) Linguistic Predictors from Facebook Postings of Substance Use Disorder Treatment Retention versus Discontinuation. *The American Journal of Drug and Alcohol Abuse Encompassing.* Vol. 48. Is. 5. P. 573–585. DOI: 10.1080/00952990.2022.2091450

Mishlanov V.A., Kadzhaya L.A., Kuznetsova Yu.M. (2020) Linguistic Markers of Emotional State of the Speech Subject (on the Problem of Automatic Monitoring of Network Communication Texts). *Medialingvistika.* Vol. 7. No. 4. P. 428–444. DOI: 10.21638/spbu22.2020.405

Petrov E.Yu., Sarkisova A.Yu. (2021) Resource of Software Platform "Polyanalyst" in Social Science and Humanities Research. *Otkrytyye dannyye — 2021: materialy foruma. Ed. by A.Yu. Sarkisova.* Tomsk: Izdatel'stvo Tomskogo gosudarstvennogo universiteta. P. 94–104.

Sboev A.G., Gudovskikh D.V., Moloshnikov I.A., Kukin K.A., Rybka R.B., Ivanov I.I., Vlasov D.S. (2013) Avtomaticheskoye vydeleniye psikholingvisticheskikh kharakteristik tekstov v ramkakh kontseptsii Big Data [Automatic selection of psycholinguistic characteristics of texts within the concept of Big Data]. *Sovremennye informacionnye tehnologii i IT-obrazovanie.* No. 9. P. 433–438.

Shchekotin E.V., Goiko V.L., Myagkov M.G., Dunaeva D.O. (2021) Assessment of Quality of Life in Regions of Russia Based on Social Media Data. *Journal of Eurasian Studies.* Vol. 12. No. 2. DOI: 10.1177/18793665211034185

Дата поступления/Received: 01.03.2023