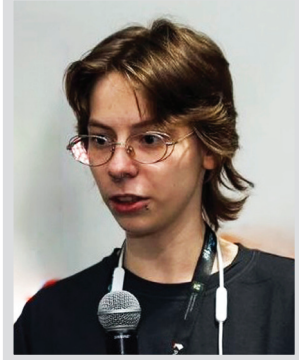
УДК 311.2
DOI: 10.24412/1998-5533-2024-4-300-303

Метод парсинга текстовых данных и его потенциал для тематического анализа как инструмент разведывательного анализа*

Мальцева А.В.
Доктор социологических наук, профессор, доцент
кафедры социального анализа и математических методов в социологии
Санкт-Петербургского государственного университета

Гуриева С.Д.
Доктор психологических наук, профессор,
заведующая кафедрой социальной психологии
Санкт-Петербургского государственного университета



Машаро Т.С.
Лаборант-исследователь факультета психологии
Санкт-Петербургского государственного университета

Актуальность темы исследования обусловлена возрастающими запросом исследователей к разработке и апробации методологии, направленной на получение пилотных результатов на основе данных социальных сетей. Цель исследования заключается в демонстрации потенциала разработанного дизайна исследования в решении обозначенной проблемы. Для достижения цели был решен ряд исследовательских задач: разработан дизайн исследования, решающий рассматриваемую проблему; дизайн был апробирован для анализа музыкальных сообществ; данные были сопоставлены с исследовательскими и аналитическими работами в данной области. Научная и практическая значимость заключается в разработке оригинального исследовательского дизайна, пригодного для решения рассматриваемой методической проблемы. Научная новизна результата подтверждается отсутствием единых общепризнанных подходов в решении обозначаемой методической проблемы, которые могли бы быть реплицированы ограниченными исследовательскими ресурсами.

Ключевые слова: *текстовые данные, парсинг, тематическое кодирование, искусственный интеллект, ВКонтакте, анализ естественного языка, онлайн сообщества*

Для цитирования: *Мальцева А.В., Гуриева С.Д., Машаро Т.С. Метод парсинга текстовых данных и его потенциал для тематического анализа как инструмент разведывательного анализа // Вестник экономики, права и социологии. 2024. № 4. С. 300–303. DOI: 10.24412/1998-5533-2024-4-300-303.*

* Работа выполнена при поддержке СПбГУ, шифр проекта 124032600013-2.

Современные проблемы в социологии требуют развития нетривиальных для дисциплины методов сбора и анализа данных, что вызвано ростом объема эмпирического материала, с которым сталкивается социолог, о чем пишут С. Ткач и др. [1]. В работе С. Ткача с соавторами отмечается, что это обусловлено трансформацией социального пространства, которое обретает гибридные формы при его преломлении в цифровых социальных сетях. Среди актуальных направлений работы в данной проблематике исследователи выделяют специфично работу с текстовыми данными и их парсинг [2]. Под парсингом данных понимается алгоритмизированный сбор интернет-данных, выполняемый в соответствии с заранее написанными командами. Команды могут включать в себя переменные, описывающие значимые для исследователя характеристики требуемых данных: их объем, адрес размещения и др. [3]. Собранная посредством парсинга текстовых данных информация может быть структурирована и проанализирована в относительно сжатые сроки, что позволит обнаружить неочевидные для исследователей паттерны в тексте, выделить основные тематические и семантические коды [4]. На парсинг текстовых данных возлагаются значительные ожидания, как на релевантный метод сбора социологических данных [5]. Так как он не предполагает активного взаимодействия с пользователем, с этим методом также связываются надежды на преодоление недоверия опросам со стороны респондентов – предполагается, что результаты, которые иначе могли бы быть получены в опросе, могут быть агрегированы в комментариях пользователей в социальных сетях [6].

Цель данной статьи – представить возможный дизайн социологического исследования, в котором получаются разведывательные (пилотные) данные о тематической области музыкальных предпочтений пользователей социальной сети ВКонтакте. Предлагаемый дизайн обладает рядом преимуществ в простоте и оперативности его воспроизведения, что позволяет достаточно быстро получить пилотное представление исследователю об изучаемом объекте.

Дизайн исследования и его обоснование

Исследование включало в себя 3 этапа: онлайн-опрос респондентов, парсинг и визуализацию данных. Визуализация данных проводилась с целью показать возможности, открывающиеся при анализе данных, полученных в рамках предыдущего этапа.

Во время первого этапа сбора данных был проведен опрос молодых людей в возрасте от 18 до 35 лет в социальной сети ВКонтакте. В опросе участвовало 200 чел. Респондентам было предложено назвать 15 сообществ на музыкальную тематику, которые, по их мнению, являются самыми популярными среди молодежи. Респонденты были выбраны случайным образом из пользователей социальной сети в разделе: «Могут быть Вам знакомы». Каждому респонденту задавался вопрос: «Какие, по Вашему мнению, 15 музыкальных групп в социальной сети ВКонтакте, являются наиболее популярными?». Также уточнялось: «Укажите ссылки на паблики, пожалуйста». Данный этап будет релевантным и для других тематик при корректировке метода отбора выборочной совокупности. К примеру, если речь идет о какой-то более узкой теме, в опросе могут принять участие эксперты, которые могут предоставить данные о тематических сообществах социальной сети ВКонтакте. Соответственно, процедура отбора респондентов будет неслучайной, будут использованы методы экспертного отбора.

Ответы были систематизированы в таблицу *Excel*. В таблице содержались несколько столбцов: *ID* сообщества, ссылка на него, характеристика респондента. Характеристика респондента включала в себя: пол, возраст, место работы/направление учебы. После объединения результатов сбора данных были удалены дубликаты, а данные о характеристиках респондентов сохранены.

Следующим этапом стала выгрузка количества подписчиков каждого сообщества с использованием метода *groups.getById* (здесь и далее – программного интерфейса ВКонтакте), после чего группы были отсортированы по убыванию числа подписчиков. Далее, с помощью метода *wall.get* были выгружены данные о 500 постах каждого сообщества. В результате была получена информация о постах, включая записи, количество лайков, комментариев и репостов, а также приложенные к постам фотографии, видео, ссылки, иные медиафайлы. Объект, получаемый в ответ на запрос *wall.get*, включал в себя определенные характеристики постов, такие как *artist* (исполнитель), *title* (название), *main_artists name* (имя основного исполнителя), *featured_artists name* (имя соисполнителя). *Title* и *text* были скопированы в отдельные файлы *Excel*. Для каждого из файлов было создано облако часто встречающихся слов при помощи библиотек *wordcloud*, *matplotlib* и *nltk* языка *Python*. Предобработка текста постов для последующего анализа включала в себя приведение текста к нижнему регистру, удаление пунктуации, а также тензорных слов (союзов, предлогов, местоимений и др.). Для предобработки использовалась библиотека *nltk*. Построение облаков слов происходило при помощи библиотеки *wordcloud*. Для графической отрисовки изображений использовалась библиотека matplotlib. Полученные облака слов учитывали частоту встречаемости слова в исходном корпусе, а сами слова в облаке слов были тем больше, чем более встречаемым было слово.

Результаты

Рассмотрим облака слов, полученные для некоторых из характеристик групп.

На рисунке 1 представлено облако часто встречающихся слов из столбца *text*, полученного методом

Рис. 1. Облако слов, полученных из столбца «text»

wall.get. Оно демонстрирует ключевые слова, которые часто используются в записях, опубликованных в сообществах *ВКонтакте*. Столбец *text* содержит текст записи, которая была опубликована в сообществе или на странице пользователя. Это может включать в себя комментарии, описания фотографий, видео, аудиозаписей, ссылок и других типов контента, которые были размещены на стене сообщества или профиля. Текст записей может быть написан на разных языках и содержать различные элементы, такие как эмодзи, смайлики, хештеги и ссылки. В данном случае самое часто встречающееся слово «комментарии» отражает актуальность и важность комментариев для сообществ и участников, а слово «сегодня» может указывать на актуальность информации или текущие события. Слово «просто» может использоваться для подчеркивания простоты или естественности сообщений. Другие ключевые слова, такие как «реально», «серьезно», «песня», «фото», «реп» и так далее, могут отражать тематику и интересы сообществ, а также стиль общения их участников.

В данном случае, если рассматривать столбец *title* как хранилище исполнителей музыкальных произведений, то облако часто встречающихся слов, полученных с помощью метода *wall.get*, может показать следующее (рис. 2). Самое часто встречающееся слово в данном столбце может быть «*Oxxxymiron*», что указывает на то, что этот исполнитель является наиболее популярным или часто упоминаемым в сообществах *ВКонтакте*, относящихся к молодежной музыкальной тематике. На втором месте может стоять «*Og Buda*», что говорит о том, что этот исполнитель также имеет значительное количество

Рис. 2. Облако слов, полученных из столбца «title»

упоминаний и, возможно, популярности в данных сообществах. На третьем месте может находиться «*Alexander Komarov*», что также указывает на его значимость и упоминаемость в контексте молодежных музыкальных сообществ *ВКонтакте*.

Заключение и дискуссия

Полученные результаты относительно музыкальных вкусов могут быть интерпретированы с опорой на выводы, полученные авторами в аналогичных работах. В исследовании М. Кашиной и коллег упоминается ряд исполнителей, отмеченных в облаке слов, в частности, исполнители лейбла *Black Star* (в частности, Anna Asti и др.). Они свидетельствуют о возрастной популярности *k-pop* музыки среди молодежи, упоминание о которой также есть в облаке слов [7]. К сходным выводам приходят Н.Г. Тагильцева и М.Н. Курлапов [8], а также А.В. Захваткин и Е.Ю. Темникова [9].

Также релевантным будет сопоставление полученных результатов с официальными статистическими сводками, публикуемыми социальной сетью ВКонтакте. Здесь однако нужно указать на ограничение такого сопоставления. *ВКонтакте* публикует данные о наиболее прослушиваемых исполнителях – множество, которое может не совпадать с наиболее популярными исполнителями. Как пишет Дж. Фиске, популярность современного автора большей частью может обеспечиваться не вниманием к его творчеству, а интересом к его личной жизни [10]. Также речь идет о популярности среди всех пользователей, а не только молодежи. Тем не менее самостоятельно такая валидация может пониматься как косвенная. Авторы, а также лейблы, указанные в облаке слов, встречаются в официальном рейтинге, полученном на основе внутренней продуктовой аналитики ВКонтакте за 2023 г. [11]: в основном это касается принадлежности значительной части авторов как в нашем исследовании к жанрам рэп и хип-хоп, а также в принадлежности многих авторов к лейблу *Black Star*.

Совокупно это позволяет говорить об удачности описываемого дизайна исследования с точки зрения получения содержательно значимых пилотных результатов на большом изначальном массиве данных с использованием небольших трудовых затрат.

Литература:

1. Ткач С., Воробьева П.Д., Русакова М.М. Опыт реализации дискурс-анализа и концептуального картирования сообществ здорового питания

// Социология: методология, методы, математическое моделирование. 2023. № 56. С. 143–172. DOI: 10.19181/4m.2023.32.1.4.
2. Kaushik A., Naithani S. A comprehensive study of text mining approach // International Journal of Computer Science and Network Security (IJCSNS). 2016. Vol. 16. №. 2. Art. 69.
3. Miller S. Fox H., Ramshaw L., Weischedel R. A novel use of statistical parsing to extract information from text // 1st Meeting of the North American Chapter of the Association for Computational Linguistics. 2000. P. 226–233.
4. Radovanović M., Ivanović M. Text mining: Approaches and applications // Novi Sad J. Math. 2008. Vol. 38. №. 3. P. 227–234.
5. Alwidian S.A., Bani-Salameh H.A., Alslaity A.N. Text data mining: a proposed framework and future perspectives // International Journal of Business Information Systems. 2015. Vol. 18. №. 2. P. 127–140.
6. Splichal S. In data we (don't) trust: The public adrift in data-driven public opinion models // Big Data & Society. 2022. Vol. 9. №. 1. Art. 20539517221097319.
7. Белая Е.К., Кашина М.А. K-pop, социальные сети, гендерные представления: проблемы презентации и восприятия (на примере творчества группы BTS) // Управленческое консультирование. 2022. № 11(167). С. 67–85.
8. Тагильцева Н.Г., Курлапов М.Н. Музыкальные предпочтения студентов негуманитарных специальностей технического университета // Педагогическое образование в России. 2024. № 2. С. 232–238.
9. Захваткин А.В., Темникова Е Ю. Музыкальные предпочтения молодежи как индикатор их психоэмоционального состояния // Ученые записки НТГСПИ. Серия: Педагогика и психология. 2023. № 4. С. 84–93.
10. Fiske J. The cultural economy of fandom // The adoring audience. – Routledge, 2002. P. 30–49.
11. VK Музыка рассказывает, кого слушали в 2023 г. // ВКонтакте. URL: https://vk.com/press/music-2023 (дата обращения: 06.10.2023).

Text Data Parsing Method and its Potential for Thematic Analysis as a Tool for Exploratory Research

Maltseva A.V., Gurieva S.D., Masharo T.S.
Saint Petersburg State University

The relevance of the research topic is due to the increasing demand of researchers for the development and testing of a methodology aimed at obtaining pilot results based on social network data. The purpose of the study is to demonstrate the potential of the developed research design in solving the designated problem. To achieve the goal, a number of research tasks were solved: a research design was developed that solves the problem under consideration; the design was tested for the analysis of musical communities; the data were compared with research and analytical works in this area. Scientific and practical significance lies in the development of an original research design suitable for solving the methodological problem under consideration. The scientific novelty of the result is confirmed by the absence of uniform generally accepted approaches to solving the designated methodological problem that could be replicated by limited research resources.

Keywords: text data, parsing, topic coding, artificial intelligence, VKontakte, natural language analysis, online communities