DOI 10.31029/vestdnc96/16

УДК 004.942

ПРОГРАММНАЯ ОБРАБОТКА ТЕКСТОВЫХ ДАННЫХ В СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Р. Г. Мифтахова, ORCID: 0009-0007-4915-8887
Уфимский университет науки и технологий, Уфа, Россия

SOFTWARE PROCESSING OF TEXT DATA IN SOCIOLOGICAL RESEARCHES

R. G. Miftakhova, **ORCID:** 0009-0007-4915-8887
Ufa University of Science and Technology, Ufa, Russia

Аннотация. Появление больших данных и вычислительных мощностей открыло новые возможности для использования крупномасштабных текстовых источников в социологических исследованиях. Недавние работы в области социологии культуры, науки, а также экономической социологии показали, как квантитативный майнинг текста может быть использован для построения и проверки социологических теорий. В данной статье рассмотрены квантитативные методы, используемые в современных исследованиях: семантический и нейросетевой анализ, языковые модели, машинное обучение и то, как они могут помочь социологам в решении различных аналитических задач.

Abstract.The emerge of big data and computing power have opened up new opportunities for the use of large-scale textual sources in sociological researches. Recent works in the field of sociology of culture, science, and economic sociology have shown how quantitative text mining can be used to construct and verify sociological theories. This article discusses quantitative methods used in modern researches: semantic and neural network analysis, language models, machine learning and how they can help sociologists solve various analytical problems.

Ключевые слова: модель языка, машинное обучение, нейросеть, квантитативная лингвистика, прикладная лингвистика.

Keywords: language model, machine learning, neural network, quantitative linguistics, applied linguistics.

Распространение больших данных и цифровизация социальной жизни привели к созданию больших объемов данных, из которых можно получить информацию в дополнение к той, которая извлекается традиционными методами.

По мере доступности данных и развития методологии на пересечении нескольких социальных научных дисциплин стали появляться квантитативные социальные науки, отвечающие на социальные вопросы, опираясь на новые источники данных и методы обработки [1]. Программный анализ текста особенно хорошо подходит для использования вычислительных методов. Социология имеет долгую историю использования компьютерного анализа как в количественных, так и в качественных исследовательских традициях. Этим объясняется то, что заметные методологические инновации были сделаны в социологических исследованиях, связанных с анализом больших объемов текстовых данных. Социологи используют новые квантитативные подходы для изучения различных явлений, включая дискурс политической элиты, риторику онлайн-сообществ [2], социальные группы и движения [3], разнообразные способы фрейминга событий в медиа [4], а также характер публичных дебатов вокруг различных общественных проблем [5]. Компьютерный анализ текста поддерживает как исследования, связанные с эволюцией культурных смыслов на макро-уровне [6], так и выводы о культурных ценностях на микроуровне [7]. Наконец, компьютерные методы помогают понять внутренние механизмы науки [8] и поддерживают новые методологические инструменты [9].

В данной статье представлены исследования, использующие возможности вычислительного анализа в построении и тестировании социологической теории. Особую роль играют методы словарей. Они количественно оценивают наличие слов в текстовых данных, связанных с интересующими концепциями, такими как негативная или позитивная речь, популизм. Не менее значимы семантические и нейросетевые методы анализа текста, их роль – облегчить идентификацию действий или действующих лиц в тексте. Эффективными являются методы кластеризации текста «без учителя» для исследователь-

ского анализа путем группировки текстов на основе скрытых тем, которые они содержат, и методов контролируемой классификации для больших объемов текстов. Все эти подходы позволяют численно представлять текст и выявлять сложные смыслы в нем.

Увеличение числа приложений, использующих квантитативный анализ текста в социологии, привело к обсуждениям надежности больших текстовых источников, практическим проблемам сбора и анализа текстовых данных, а также последствиям внедрения индуктивных вычислительных методов в инструментарий социологии [10]. Поскольку такие обеспокоенности являются неотъемлемыми для любого приложения, использующего большие текстовые данные и вычислительные методы, в данной работе также излагаются основные аргументы, выдвинутые в этих дискуссиях.

Остановимся на способах словарных методов. В приложениях анализа текста словари обозначают списки слов, связанных с концепцией интереса. Словарные методы автоматически идентифицируют слова из словаря в тексте и, следовательно, особенно эффективны при определении и количественной оценке наличия интересующей концепции, например, позитивной или оскорбительной лексики, в больших текстовых корпусах. При этом можно опираться на эмпирически проверенные существующие словари, создавать свои собственные словари, адаптированные к конкретному вопросу исследования или комбинировать эти ресурсы. Существующие словари особенно эффективны при поиске концепций, которые были идентифицированы в предыдущих исследованиях, например, слов, обозначающих положительные или отрицательные эмоции. Некоторые из таких ресурсов были обширно подвергнуты эмпирической проверке. Например, “General Inquirer” содержит многочисленные категории словарей, указывающие на несколько эмоций, политические концепции и аспекты теории мотивации. Аналогично, словарь программного обеспечения “Linguistic Inquiry and Word Count (LIWC)” содержит 82 категории, касающиеся различных психологических процессов. В исследовании Голдера и Мейси [11] используется словарь LIWC для идентификации положительных и отрицательных слов в 509 млн сообщений от 2,4 млн пользователей в 84 странах. Авторы показывают, что короткие сообщения в социальных сетях в целом отмечены как положительные в выходные дни и утром, что подтверждает гипотезу о том, что тексты в социальных медиа отражают биологические ритмы. Бейл и др. [5] используют словарь LIWC для идентификации элементов аффективного и рационального стилей речи в сообщениях в социальных медиа. Анализ контента, размещенного в социальных медиа 92 организациями за 1,5 года, поддерживает гипотезу о существовании циклов эмоциональной и рациональной речи в онлайн-дискуссиях [5]. При таком анализе один из подходов заключается в преобразовании текстового корпуса в матрицу документ-термин (DTM), где каждая строка матрицы представляет собой текст, каждый столбец - слово, а каждое поле – количество раз, которое каждое слово появляется в каждом тексте. Одна строка такой матрицы документ-термин содержит подсчеты всех слов в отдельном тексте, фактически преобразуя его в числовое векторное представление. В рамках данной статьи автором был создан корпус в сорок тысяч слов соответственно из текстов англоязычных и русскоязычных средств массовой информации за последние два года. Некоторые модели были протестированы на этих корпусах. Самыми частотными словами в русском корпусе оказались слова «Россия, Украина, туристы», в англоязычном – “vehicle, US, Russia, Ukraine”. Однако относительно небольшой размер корпуса не позволяет получить более достоверные результаты, что не преуменьшает функциональности использованной модели.

В некоторых исследованиях возникает необходимость учитывать лингвистический контекст каждого слова, т.е. несколько слов, предшествующих и следующих за ним. В таких приложениях можно вычислить условную вероятность того, что слово появится в определенной позиции в тексте [12]. Макмахан и Эванс [8] полагаются на аналогичную модель генерации языка для выявления неоднозначности в научных исследованиях. Эта модель позволяет авторам оценить вероятность того, что слово приобретет определенное значение в зависимости от его лингвистического контекста: чем больше синонимов слова встречается в том же контексте в теле соответствующих текстов, тем более неоднозначно значение слова.

Тексты в нейронных сетях преобразовываются в числа с помощью так называемых распределенных текстовых представлений. Эти представления – модели вложения слов – представляют слова в виде векторов в плотном пространстве высокой размерности, где слова, встречающиеся в схожих контекстах в анализируемом корпусе текстов, расположены близко друг к другу в пространстве [13]. По сравнению с

другими методами, модели вложения слов особенно подходят для представления многоаспектных ассоциаций между словами и улавливания культурных нюансов значения в больших, сложных текстовых корпусах [6]. Существующие доказательства указывают на то, что модели вложения слов отображают значение в текстах таким образом, который можно сравнить с представлениями человека.

Исследование Козловского показывает, как модели встраивания слов могут улавливать сложные культурные категории и связи между ними. Авторы представляют слова, найденные в миллионах книг, опубликованных за столетие, в виде векторов размерности 300 и используют эти векторы для построения культурных измерений социального класса, таких как богатство или статус, и отслеживают как со временем меняется их положение в пространстве языковых векторов. Такой подход позволяет проводить очень детальный анализ динамики культурных измерений класса: например, в то время как образование ассоциировалось с богатством через сложности в 1900-х гг., к 1990-м эта ассоциация стала прямой [6]. Другие исследования использовали мощь моделей встраивания слов для отслеживания эволюции гендерных стереотипов с использованием большого корпуса печатных СМИ, а также исследования связи между общественным дискурсом вокруг пандемии коронавируса и склонностью жителей к социальной дистанции в различных сообществах. Автором статьи был создан и проанализирован корпус СМИ за период пандемии и выявлено, что семантическое поле слова «коронавирус» в русскоязычных корпусах ассоциировалось со словами «самоизоляция», «выплаты и компенсации», «поддержка бизнеса», тогда как в англоязычных корпусах семантика «коронавируса» связывалась с «вакцинами», «нехваткой койкомест», «лечением» [14, с. 305]. Классические же методы анализа текста не позволяют выявить такие тонкие семантические различия, связанные с разной социальной средой.

Перспективные разработки в методологии текстового майнинга, которые могут быть интересны социологам, включают улучшения в создании и анализе словарей, новые инструменты для семантического и сетевого анализа, представления текста и моделей машинного обучения, а также использование методов глубокого обучения. Создание пользовательских словарей – довольно сложная задача, исследователям приходится изучать большое количество текстов, чтобы получить достаточно широкий набор слов, соответствующих их концепциям интереса. В недавних работах представлены методы, основанные на концепции ключевых слов из вычислительной лингвистики, с целью облегчить поиск интересующих слов в неструктурированном тексте и упростить автоматическое создание словарей из вручную маркированных данных для нескольких категорий [7]. Эти программные реализации выполняют более точный анализ словарей: они могут учитывать слова, которые корректируют или изменяют валентность интересующего слова. Такие методы, выполняющие автоматический семантический и сетевой анализ, требуют значительных вычислительных ресурсов.

Что касается цифрового представления языка, оно позволяет исследовать эволюцию языка и культурные значения в больших текстовых данных.

Увеличение популярности моделей вложения слов привело к разработке моделей на основе различных текстовых ресурсов. Например, были собраны языковые модели на 157 различных языках.

Перспективной представляется технология переноса машинного обучения. Языковые модели, созданные в одном домене, можно использовать в других доменах. В последние несколько лет новые разработки в области глубокого обучения значительно улучшили качество цифровых представлений слов. Подобно тому, как мозг человека обрабатывает информацию, алгоритмы глубокого обучения итеративно преобразуют начальную текстовую информацию в ее более абстрактные цифровые представления. Такие представления слов и словосочетаний способны оценивать многообразие контекстов и лучше улавливать скрытые смыслы. При выполнении задач кластеризации они указывают на слова, вокруг которых должны формироваться темы. В таких решениях комбинируются неконтролируемые модели выявления темы текста и контролируемые модели их классификации, что может значительно приблизить эффективность и качество машинного обучения к анализу текста социологом-экспертом. Хотя на данный момент нет широко известных приложений в социологии, использующих эти методы при анализе текста, но, учитывая их растущую популярность в других областях, отличные результаты и увеличивающуюся доступность инструментов, эти методы, скорее всего, заменят более простые реализации машинного обучения.

Особо стоит отметить данные, сгенерированные в онлайн-среде. Они позволяют исследователям захватить множество следов реального поведения людей в реальных ситуациях и в реальном времени, предоставляя возможность получить знания или идеи, которым было бы сложно появиться при использовании традиционных методов социологических исследований. Такие данные могут быть получены с различных платформ (например, через API доступ, предоставляемый Twitter) или собраны с использованием пользовательских скриптов “crawlers”, которые автоматически просматривают веб страницы и скачивают данные.

В дополнение к вышеобозначенным преимуществам сбор больших текстовых данных позволяет преодолеть внутреннюю противоречивость других методов сбора данных [11, с. 241], а также проблемы памяти и предвзятости в ответах. Тем не менее эти данные чаще всего не создаются с конкретными исследовательскими целями. Поэтому они могут быть лишены необходимой для исследования информации, например, уровень образования авторов отдельных текстов, их социально-экономический статус, индивидуальные психологические особенности; контекст создания текста. Так, корпус Google Books, в котором миллионы книг, часто лишен надежных метаданных. Решением для таких проблем может стать объединение различных больших источников данных и их интегрирование с традиционными социологическими источниками, например, данными из опросов. Кроме того, при сборе текстовых данных полезно руководствоваться теорией, учитывать дополнительную информацию, необходимую для ответа на конкретный исследовательский вопрос.

Большие объемы текста в социологических исследованиях – описательные или гипотезированные – необходимо преобразовать в управляемый ввод. Текст существенно отличается от традиционных количественных данных, используемых социологами. Текстовые данные являются высокоразмерными и разреженными. Что касается высокоразмерности, то даже небольшой объем текстов может содержать тысячи уникальных слов: захват позиции каждого слова в его контексте в каждом тексте требует представления данных в высокоразмерной форме. А разреженность характеризуется тем, что каждый текст может содержать только небольшое количество слов, в то время как количество слов во всех рассматриваемых текстах может быть довольно большим. Хотя современные алгоритмы текстового майнинга способны обрабатывать разреженные высокоразмерные данные, включение изобилия текстовой информации в анализ не обязательно улучшает эффективность методов текстового майнинга. Напротив, работа с разреженными данными и слишком многими переменными может легко привести к переобучению модели.

Следовательно, нужно уменьшить размерность и разреженность данных. Для уменьшения количества признаков и сокращения избыточной обработки целесообразно удалять редко встречающиеся слова и использовать автоматические методы, оставляя только те признаки, которые полезны для исследования. Можно дополнительно упростить представления текста, не учитывать контекст, в котором появляется каждое слово, так называемый метод «мешка слов», или использовать векторные представления слов. При работе с методами машинного обучения важно не допустить переобучения модели на данных, используемых для ее обучения. Наличие переобучения можно проверить, разделив набор данных на несколько подмножеств – одно, на котором модель «обучается», и другое, на котором она «проверяется». Если модель хорошо справляется с первым подмножеством, но плохо – со вторым, вероятно, имеет место переобучение. Помимо правильной подготовки данных, можно выбирать алгоритмы, которые менее подвержены переобучению или искать более сложные решения для борьбы с этой проблемой.

На современном этапе обработки текста автоматические методы служат всего лишь инструментами для улучшения и дополнения анализа, выполненного человеком; их эффективность должна быть подтверждена путем сравнения с эталоном. Этап валидации зависит от подхода и исследовательского вопроса: он может варьироваться от сравнения эффективности автоматических методов с золотым стандартом экспертного ручного кодирования – при контролируемой классификации текста - до глубокого анализа выборки текстов исследователями – при неконтролируемой кластеризации текста. Кроме того, в последнее время в работах проводится более тщательная проверка измеряемых в тексте концепций сравнительно с аналогичными измерениями, полученными с помощью методов опроса.

Наконец, возникают новые этические вопросы относительно сбора, использования и хранения больших источников данных, собранных онлайн. В отличие от участников опросов, интервью или экс-

периментальных исследований, люди, занимающиеся повседневной онлайн-деятельностью, не предоставляют свое согласие исследователям так явно. Более того, данные, создаваемые онлайн, могут легко связываться между онлайн-платформами и источниками, если они недостаточно анонимны, возможно, позволяя идентифицировать людей без их согласия.

ЛИТЕРАТУРА

1. *Edelmann Achim, Wolff Tom, Montagne Danielle, Bail Christopher A.* Computational social science and sociology // Annu. Rev. Sociol. 2020. Vol. 46. P. 61–81.

2. *Davidson Thomas, Warmsley Dana, Macy Michael, Weber Ingmar*. Automated hate speech detection and the problem of offensive language // Proceedings of the International AAAI Conference on Web and Social Media, 2017. Vol. 11. P. 512–515.

3. *Almquist Zack W., Bagozzi Benjamin E.* Using radical environmentalist texts to uncover network structure and network features // Socio. Methods Res. 2019. Vol. 48 (4). P. 905–960.

4. *Bastin Gilles, Bouchet-Valat Milan.* Media corpora, text mining, and the sociological imagination – a free software text mining approach to the framing of Julian assange by three news agencies using R.TeMiS // Bulletin of Sociological Methodology / Bulletin de M´ethodologie Sociologique. 2014. Vol. 122(1). P. 5–25.

5. *Bail Christopher A., Brown Taylor W., Mann Marcus*. Channeling hearts and minds: advocacy organizations, cognitive-emotional currents, and public conversation // Am. Socio. Rev. 2017. Vol. 82 (6). P. 1188–1213.

6. *Kozlowski Austin C., Taddy Matt, Evans James A.* The geometry of culture: analyzing the meanings of class through word embeddings // Am. Socio. Rev. 2019. Vol. 84 (5). P. 905–949.

7. *Macanovic Ana, Przepiorka Wojtek.* The Moral Embeddedness of Cryptomarkets: Text Mining Feedbacks on Economic Exchanges in the Darknet. Department of Sociology / ICS, Utrecht University, The Netherlands. Unpublished manuscript. 2021.

8. *McMahan Peter, Evans James*. Ambiguity and engagement // Am. J. Sociol. 2018. Vol. 124 (3). P. 860–912.

9. *Nardulli Peter F., Althaus Scott L., Hayes Matthew.* A progressive supervised-learning approach to generating rich civil strife data // Socio. Methodol. 2015. Vol. 45 (1). P. 148–183.

10. *Goldberg Amir, Srivastava Sameer B., Govind Manian V., Monroe William, Potts Christopher.* Fitting in or standing out? The tradeoffs of structural and cultural embeddedness // Am. Socio. Rev. 2016. Vol. 81 (6). P. 1190–1222.

11. *Golder Scott A., Macy Michael W.* Digital footprints: opportunities and challenges for online social research // Annu. Rev. Sociol. Vol. 40 (1). P. 129–152.

12. *Jurafsky Daniel, Martin James H.* Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson / Prentice Hall, Upper Saddle River, NJ, 2009. 1044 p.

13. *Mikolov Tomas, Wen-tau Yih, Zweig Geoffrey.* Linguistic regularities in continuous space word representations // In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013. P. 746–751.

14. *Мифтахова Р.Г.* Выявление оттенков смысла в текстах СМИ // Языки в диалоге культур: проблемы многоязычия в полиэтническом пространстве : Материалы III Всероссийской научно-практической конференции с международным участием, посвященной 75-летию Победы в Великой Отечественной войне / отв. ред. *Р.А. Газизов*. Уфа, 2020. С. 300–303.

Поступила в редакцию 15.10.2024 г.
Принята к печати 26.03.2025 г.

* * *

Мифтахова Рамиля Габдулхаевна, *кандидат филологических наук, доцент кафедры лингводидактики и переводоведения, Высшая школа зарубежной филологии, лингвистики и перевода, Институт гуманитарных и социальных наук, Уфимский университет науки и технологий; e-mail: miftahovar@yandex.ru*

Ramilya G. Miftakhova, *Candidate of Philology, associate professor of the Department of Linguodidactics and Translation Studies; Higher School of Foreign Philology, Linguistics and Translation, Institute of Humanities and Social Sciences, Ufa University of Science and Technology; e-mail: miftahovar@yandex.ru*