

Опыт применения больших языковых моделей для анализа количественных социологических данных

Ашихмин Евгений Георгиевич

Аспирант

ORCID: 0000-0001-9193-4535, e-mail: e.ashikhmin@icloud.com

Левченко Валерий Витальевич

Д-р психол. наук, зав. каф. социологии и политологии

ORCID: 0000-0002-7627-9162, e-mail: levv66@mail.ru

Селеткова Гюзель Ильясовна

Ст. преп. каф. социологии и политологии

ORCID: 0000-0003-3402-3473, e-mail: guzal.ka@mail.ru

Пермский национальный исследовательский политехнический университет, г. Пермь

Аннотация

В работе рассматриваются возможности и ограничения применения больших языковых моделей для анализа количественных данных в социологических исследованиях. Также уделяется внимание акторно-сетевой теории, согласно которой нейронные сети выступают как активные участники социального взаимодействия. Отмечается, что применение LLM (англ. large language model – большая языковая модель) можно рассматривать как инновационный процесс в сфере прикладных социологических исследований. Демонстрируются примеры применения LLM для количественных методов анализа на основе массива данных опроса, взятого из открытых источников. На практических примерах показано, как большие языковые модели могут использоваться для построения частотных и сводных таблиц, расчета средних значений и для проведения корреляционного анализа. Применение LLM рассматривается как инновационный процесс, способствующий развитию новых методологических подходов. Анализируются примеры использования LLM в социологии и подчеркивается необходимость формирования инновационной культуры и развития методологических подходов для проверки и коррекции результатов. Кроме того, авторы подчеркивают важность интерпретации результатов, полученных с помощью больших языковых моделей, в контексте социологической теории и практики. Также обсуждается роль LLM в расширении возможностей социологических исследований, особенно в области анализа больших данных и обнаружения скрытых паттернов. Наконец, авторы предлагают пути будущих исследований в сфере применения LLM в социологии, включая разработку новых методов и инструментов для интеграции больших языковых моделей в социологические исследования.

Ключевые слова

Большие языковые модели, LLM, цифровые инструменты, анализ количественных данных, методы социологических исследований, цифровая трансформация, частотные распределения, сводные таблицы, расчет средних значений, расчет корреляции

Для цитирования: Ашихмин Е.Г., Левченко В.В., Селеткова Г.И. Опыт применения больших языковых моделей для анализа количественных социологических данных // Вестник университета. 2024. № 11. С. 205–215.

Experience in applying large language models to analyse quantitative sociological data

Evgenii G. Ashikhmin

Postgraduate Student

ORCID: 0000-0001-9193-4535, e-mail: e.ashikhmin@icloud.com

Valery V. Levchenko

Dr. Sci. (Psy.), Head at the Sociology and Political Science Department

ORCID: 0000-0002-7627-9162, e-mail: levv66@mail.ru

Gyuzel' I. Seletkova

Senior Lecturer at the Sociology and Political Science Department

ORCID: 0000-0003-3402-3473, e-mail: guzal.ka@mail.ru

Perm National Research Polytechnic University, Perm, Russia

Abstract

The article discusses the possibilities and limitations of using large language models (hereinafter referred to as LLM) to analyse quantitative data in sociological research. Also, attention is paid to the actor-network theory, according to which neural networks act as active participants of social interaction. It is noted that the usage of the LLM can be considered as an innovative process in the field of applied sociological research. The article demonstrates examples of the LLM application for quantitative methods of analysis on the basis of a survey dataset taken from open sources. Practical examples show how the LLM can be used to construct frequency and summary tables, calculate averages and conduct correlation analysis. The application of the LLM is seen as an innovative process that promotes the development of new methodological approaches. The authors analyse examples of the LLM usage in sociology and emphasise the need to build an innovative culture and develop methodological approaches to verify and correct the results. In addition, the authors highlight the importance of interpreting the LLM results in the context of sociological theory and practice. The article also discusses the role of the LLM in empowering the sociological research, especially in the areas of analysing big data and discovering hidden patterns. Finally, the authors suggest paths for future research in the application of the LLM in sociology, including the development of new methods and tools for integrating the LLM into the sociological research.

Keywords

Large language models, digital tools, quantitative data analysis, sociological research methods, digital transformation, frequency distributions, summary tables, calculation of mean values, calculation of correlation

For citation: Ashikhmin E.G., Levchenko V.V., Seletkova G.I. (2024) Experience in applying large language models to analyse quantitative sociological data. *Vestnik universiteta*, no. 11, pp. 205–215.

ВВЕДЕНИЕ

Современное общество переходит на новую стадию технологических изменений, на которой нейронные сети (далее – нейросеть) играют ключевую роль. Э. Бриньольфссон, А. Макафи полагают, что искусственный интеллект (далее – ИИ) и нейросети могут стать самой важной технологией общего назначения современной эпохи [1]. Нейросети, в том числе и большие языковые модели, начинают применяться в различных областях: экономике, маркетинге, социологии и др. При этом достижения в области ИИ могут приводить к серьезным последствиям на рынке труда [2–4], трансформируя профессиональные компетенции специалистов разных сфер [5], в том числе и социологов, изменяя их стратегию и влияя на производительность труда. Некоторые исследователи акцентируют внимание на том, что применение ИИ, нейросетей может привести к усилению цифрового неравенства¹ и неравенства в целом [6–7].

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ

На применение нейросетей в исследованиях можно посмотреть с точки зрения акторно-сетевой теории [8], согласно которой нейросети выступают как определенный актор, влияющий на социальную действительность. Они все чаще используются для систематизации больших данных, прогнозирования результатов влияния на социум. Тем самым нейросети глубоко проникают в общество, представляя собой социотехнические системы, и в них ценности, институциональные практики и неравенство встроены в сам код, дизайн и практики использования ИИ. Специалисты по ИИ, нейросетям для обучения LLM (англ. large language model – большая языковая модель) вводят существующие данные, чтобы в итоге LLM могли принимать автономные решения, которые активно реинтегрируются обратно в социум [9–11]. Например, поддерживая или принимая решения о том, какие дороги необходимо ремонтировать, кому выплачивать социальные пособия [12–13]. С точки зрения акторно-сетевого подхода можно посмотреть, какие взаимосвязи выстраиваются между профессионалами и нейросетями, как они связаны с организационным контекстом, пространственным распределением рабочих мест и системами неравенства [14–15].

Применение LLM в анализе результатов социологических исследований может привести к изменению компетенций социолога, а именно к приобретению технических навыков работы с соответствующими инструментами и программным обеспечением, визуализации данных, а также к пониманию основных принципов машинного обучения². Также социологам важно развивать навыки интерпретации данных, полученных с помощью нейросетей, чтобы понимать, какие выводы можно сделать на основе анализа больших объемов данных. Для уменьшения искажений и ошибок в интерпретации результатов важны навыки критического мышления, умения оценивать методы и модели, используемые в нейросетевом анализе. Социологи должны быть гибкими и готовыми к обсуждению этических вопросов, связанных с применением нейросетей, а также к инновациям, поскольку новые методы анализа могут потребовать пересмотра традиционных подходов к социологическому исследованию.

Применение LLM можно рассматривать как инновационный процесс. При этом стоит отметить, что долгое время к инновациям, в том числе и к нейросетям, преобладал подход в рамках технико-технологической и экономической парадигмы, что привело к господству коммерческих показателей при оценке инновационной деятельности, включая большие языковые модели. Следствием длительного доминирования коммерческих аспектов в инновационной теории стали противоречия в практической реализации инновационных проектов и моделей LLM, такие как низкая инновационная восприимчивость и мотивация, с техническими возможностями, опережающими способности адаптации людей к изменениям, нехваткой управленческих ресурсов, что связано с несформированностью инновационной культуры [16]. Таким образом, необходимо больше внимания уделять формированию инновационной культуры, в частности культуре применения LLM в социологических исследованиях, повышению уровня информированности общества о целях и преимуществах использования больших языковых моделей как примера инноваций, а также развитию мотивации и готовности к применению LLM в социологических исследованиях с учетом этических аспектов.

¹ Казун А. Цифровой разрыв: зачем нужны ученые после появления ChatGPT? Режим доступа: <https://sociodigger.ru/articles/articles-page/cifrovoy-razryv-zachem-nuzhny-uchenye-posle-pojavlenija-chatgpt> (дата обращения: 25.07.2024).

² Малышев А. Нейросети в работе исследователя. Режим доступа: <https://sociodigger.ru/articles/articles-page/neiroseti-v-rabote-issledovatelja> (дата обращения: 25.07.2024).

С одной стороны, традиционные методы анализа результатов социологического исследования могут иметь ограничения в точности, эффективности и скорости обработки больших объемов данных и LLM способны помочь автоматизировать и улучшить процесс их обработки, способствуя повышению эффективности анализа данных, возможности обрабатывать их большие объемы, выявлению трендов и ключевой информации, повышению точности и глубины анализа, выявлению скрытых связей, что в конечном итоге может привести к более качественным выводам и решениям на основе полученных данных. Однако, с другой стороны, при применении больших языковых моделей в социологическом исследовании возникает ряд проблем, требующих дальнейшего изучения:

- точность интерпретации (как точно модели LLM могут интерпретировать и анализировать ответы на вопросы анкеты? Существует ли риск искажения результатов из-за особенностей работы моделей?);
- обработка неструктурированных данных (как модели LLM могут обрабатывать неструктурированные данные, такие как открытые ответы на вопросы, сравнительно с традиционными методами?);
- приватность данных (как обеспечить конфиденциальность и безопасность данных респондентов при использовании моделей LLM для анализа результатов анкетирования?);
- сравнение с традиционными методами (как сравнивать эффективность и точность использования моделей LLM с традиционными методами статистического анализа результатов анкетирования?).

Также возможно, что большие языковые модели умеют принимать решения на основе обучающих данных, которые могут содержать предвзятости или искажения, что способно повлиять на результаты анализа. Для успешного применения моделей LLM в анализе результатов исследования необходимо учитывать данные противоречия и контекст исследования, разрабатывать методы проверки и коррекции результатов, чтобы минимизировать возможные ошибки и искажения. Хотя примеры применения нейронных систем в социологии есть. Так, В.В. Круглов, М.И. Дли показали пример использования нейросетей для кластерного анализа [17], А.Ф. Шакирова – для построения индекса социальных настроений [18], Е.Е. Фомина – для получения кластеров респондентов в зависимости от их мотивации к благотворительной деятельности [19]. А.В. Мальцева, Н.Е. Шилкина и О.В. Махныткина показали возможности методов Data Mining (англ. добыча данных) в анализе больших объемов данных [20]. Д. Котов приводит пример применения ChatGPT для проведения факторного анализа³. Важно проводить дополнительные исследования и тестирования для подтверждения достоверности полученных результатов с помощью именно LLM.

Целью данной статьи является оценка потенциала применения языковых моделей (LLM) в анализе количественных данных социологического исследования.

КРАТКИЕ СВЕДЕНИЯ О НАИБОЛЕЕ РАСПРОСТРАНЕННЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ

OPENAI GPT-4O

OpenAI GPT-4o⁴, выпущенная в 2024 г., является одной из самых передовых и широко используемых больших языковых моделей на сегодняшний день. Она была разработана на основе предыдущих версий GPT с применением новейших алгоритмов машинного обучения и огромных объемов текстовых данных для тренировки.

GPT-4o демонстрирует впечатляющие возможности в генерации высококачественного текста, понимании контекста и адаптации к различным задачам. Модель способна генерировать связные и логичные тексты на самые разные темы: от новостных статей до художественной прозы и поэзии. Она также показывает высокие результаты в задачах вопросно-ответного диалога, классификации текстов, перефразирования, резюмирования и перевода между языками.

Одним из ключевых преимуществ GPT-4o является ее способность к дообучению на небольших наборах данных, что позволяет быстро адаптировать модель к специфическим предметным областям и задачам. Такую возможность предоставляют расширения – GPT (англ. GUID Partition Table – таблица разделов GUID), которые можно настроить для решения конкретных задач⁵.

³ Котов Д. Алгоритмы искусственного интеллекта в прикладных социологических исследованиях. Режим доступа: https://sociodigger.ru/articles/articles-page/algorithmy-iskusstvennogo-intellekta-v-prikladnykh-sociologicheskikh-issledovaniyakh#_ftn1 (дата обращения: 25.07.2024).

⁴ OpenAI. Hello GPT-4o. Режим доступа: <https://openai.com/index/hello-gpt-4o/> (дата обращения: 27.07.2024).

⁵ Eloundou T., Manning S., Mishkin P., Rock D. GPTs are GPTs: an early look at the labor market impact potential of large language models. Режим доступа: <https://arxiv.org/abs/2303.10130> (дата обращения: 27.07.2024).

Несмотря на свою мощь, GPT-4o не лишена некоторых ограничений. Модель может генерировать фактические ошибки, проявлять предвзятость и производить текст, не соответствующий этическим нормам. Разработчики постоянно работают над улучшением безопасности и надежности модели, но эти проблемы пока остаются актуальными.

ANTHROPIC CLAUDE 3.5

Anthropic Claude 3.5⁶, также известная как Claude 3.5 Sonnet, была представлена в 2024 г. как конкурент GPT-4o. Модель разработана с целью предоставить пользователям мощные возможности генерации текста, программного кода, сравнимые с GPT-4o.

Claude 3.5 демонстрирует высокую производительность в широком спектре задач, включая генерацию текста, вопросно-ответный диалог, классификацию текстов, анализ данных. Модель показывает значительные результаты в задачах, требующих логического рассуждения и последовательности.

Anthropic придерживается принципов ответственного развития ИИ и предоставляет подробную документацию и библиотеку различных запросов (промтов).

Несмотря на свои сильные стороны, Claude 3.5 уступает GPT-4o в некоторых аспектах, таких как разнообразие генерируемого текста и способность к быстрой адаптации, но превосходит ее в аналитических задачах. Модель также ограничена в своих возможностях обработки больших объемов данных. Для некоторых задач требуется разделять данные на несколько частей.

GOOGLE GEMINI

Google Gemini⁷, особенно в своей последней версии 1.5 Pro, представляет собой мощную альтернативу GPT-4o и Claude 3.5. Модель была выпущена в 2024 г. и быстро завоевала популярность.

Gemini 1.5 Pro демонстрирует значительные результаты в тестах, обходя конкурентов в ряде задач. Модель отличается способностью к отличному пониманию контекста, распознаванию сущностей и извлечению смысловых связей из текста. Это делает ее отличным инструментом для задач, связанных с анализом и интерпретацией данных.

Одним из уникальных преимуществ Gemini 1.5 Pro является ее способность к использованию не только текстовых, но и визуальных данных. Это позволяет модели эффективно работать с комбинированными данными, такими как изображения с подписями или диаграммы с пояснениями.

Модель все еще находится в процессе активной разработки и тестирования. Она может демонстрировать непоследовательность в результатах и требует дальнейшей оптимизации.

СРАВНЕНИЕ И ВЫВОДЫ

Сравнение GPT-4o, Claude 3.5 и Gemini 1.5 Pro показывает, что каждая из этих моделей имеет свои особенности и преимущества. GPT-4o остается лидером в генерации текста и адаптации к различным задачам, Claude 3.5 выделяется своими аналитическими возможностями, а Gemini 1.5 Pro демонстрирует возможности работы как с текстом, так и с изображениями.

Выбор конкретной модели для различных типов задач зависит от конечных потребностей пользователя: бюджет, требуемые функции, предпочтения в области безопасности и этики, а также специфика решаемых задач.

Все популярные большие языковые модели продолжают развиваться и совершенствоваться, и в ближайшем будущем мы можем ожидать дальнейшего повышения их производительности и расширения возможностей.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ АНАЛИЗ

ПРАКТИЧЕСКИЕ АСПЕКТЫ ПРИМЕНЕНИЯ LLM ДЛЯ АНАЛИЗА КОЛИЧЕСТВЕННЫХ СОЦИОЛОГИЧЕСКИХ ДАННЫХ

Применение больших языковых моделей для анализа количественных данных открывает новые возможности в социологических исследованиях. Одним из ключевых преимуществ LLM является их способность работать с неструктурированными текстовыми данными, что позволяет анализировать открытые вопросы анкет и другие нечисловые источники информации.

⁶ Anthropic. Официальный сайт. Режим доступа: <https://www.anthropic.com> (дата обращения: 28.07.2024).

⁷ Gemini. Официальный сайт. Режим доступа: <https://gemini.google.com> (дата обращения: 29.07.2024).

Использование LLM устраняет необходимость предварительного кодирования ответов респондентов, поскольку модель сама способна распознавать семантические закономерности в текстовых данных. Это упрощает процесс подготовки данных и снижает риск влияния субъективных факторов исследователя на результаты.

Кроме того, LLM могут выполнять сложные аналитические задачи, такие как кластеризация респондентов по их ответам, выявление латентных тем и факторов, построение типологий и концептуальных моделей. Это обеспечивает более глубокое понимание исследуемых явлений.

Важной особенностью LLM является возможность объяснять полученные результаты в форме связного текста, что облегчает интерпретацию данных и формулировку выводов. Большие языковые модели способны генерировать аналитические отчеты и тексты, используя научный стиль изложения.

Следует отметить, что на сегодняшний день LLM все еще имеют ограничения в работе с количественными данными и должны использоваться совместно с традиционными статистическими методами анализа. Однако развитие технологий машинного обучения позволяет прогнозировать возрастание роли больших языковых моделей в анализе статистических данных в будущем.

Для проверки возможностей анализа количественных социологических данных при помощи LLM был взят массив результатов опроса, проведенного одним из российских вузов, – особенности деятельности самозанятых граждан на современном рынке труда [25]. Объем выборки – 1 215 респондентов. Массив данных, используемый для апробации, на момент написания текста настоящей статьи находился в открытом доступе и его можно было свободно скачать.

При помощи LLM на указанном массиве данных были апробированы следующие методы количественного анализа: частотные таблицы, сводные таблицы, средние значения и корреляции.

ЧАСТОТНЫЕ ТАБЛИЦЫ

Частотные таблицы являются одним из основных инструментов описательной статистики, позволяющим сжато представить распределение ответов респондентов на вопросы анкеты. При анализе количественных данных с помощью LLM возможно автоматическое построение частотных таблиц для любого вопроса анкеты.

Входной промт для LLM может выглядеть следующим образом: «Построй частотную таблицу для вопроса № X, учитывая все варианты ответов респондентов. В таблице должны быть указаны: частота, процент, валидный процент, накопленный процент».

В результате LLM сформирует требуемую частотную таблицу, подсчитав необходимые статистики для каждого варианта ответа. Это позволяет быстро получить первичное описание распределения данных без дополнительных вычислений.

Таблица 1

Ответ LLM (табличное представление): частотная таблица

Вопрос: в каком качестве Вы работаете в настоящее время?	Частота	Процент	Валидный процент
В качестве работника предприятия (организации), занятого по найму	1 025	84,4	84,4
Руководителем предприятия (организации) с привлечением других работников на постоянной основе (работодатель)	67	5,5	5,5
В качестве работника собственного предприятия или собственного бизнеса (доходного занятия) без привлечения других работников	122	10	10
Другое	1	0,1	0,1
Всего	1 215	100	100

Составлено авторами по материалам исследования

Помимо табличного представления, LLM может генерировать текстовые интерпретации полученных частот, выделяя наиболее характерные варианты ответов, группируя близкие по смыслу ответы и т.д.

Ответ LLM (описание): «Подавляющее большинство респондентов (84,4 %) работают в качестве наемных работников предприятий или организаций. Это свидетельствует о том, что традиционная модель занятости по найму остается наиболее распространенной формой трудовых отношений. Только 5,5 % опрошенных являются работодателями, руководителями предприятий с наемными работниками. Интересно, что 10 % респондентов работают на себя, являясь индивидуальными предпринимателями или самозанятыми без привлечения наемных работников. Эта категория представляет собой сектор малого бизнеса и фрилансеров. Незначительная доля (0,1 %) респондентов указали другой вариант занятости, не вошедший в перечисленные категории. В целом данные демонстрируют преобладание наемного труда и относительно небольшую долю самозанятых и работодателей в структуре занятости населения».

Использование LLM для построения частотных таблиц избавляет исследователя от рутинных подсчетов, снижает вероятность ошибок, связанных с человеческим фактором, и ускоряет процесс первичного анализа данных.

СВОДНЫЕ ТАБЛИЦЫ

Сводные таблицы представляют собой мощный инструмент для анализа и наглядного представления многомерных данных. Построение сводных таблиц вручную может быть трудоемким и подверженным ошибкам процессом. Применение LLM позволяет автоматизировать их создание на основе неструктурированных данных анкет.

Для получения сводной таблицы достаточно сформулировать соответствующий промт для LLM, указав переменные, по которым необходимо произвести группировку данных, а также меры, которые требуется рассчитать в таблице (частоты, средние, суммы и т.д.).

Промт может выглядеть следующим образом: «Построй сводную таблицу по вопросам № X и № Y. Подсчитывай количество значений, без процентов».

В ответ LLM сгенерирует требуемую сводную таблицу в соответствии с запросом, автоматически выполнив все необходимые вычисления статистических показателей. Указанная в примере таблица построена по вопросам:

- в какой мере Вы удовлетворены содержанием Вашего труда по основному месту работы?
- оформлена ли Ваша основная работа (занятие) официально трудовым соглашением (договором, контрактом) между Вами и работодателем (заказчиком) или уполномоченными органами?

Таблица 2

Ответ LLM (табличное представление): сводная таблица

Ответ на вопрос	Затрудняюсь ответить	Полностью удовлетворен(а)	Скорее не удовлетворен(а)	Скорее удовлетворен(а)	Совсем не удовлетворен(а)	Итого
Оформлена	35	245	177	555	32	1 044
Не оформлена	5	44	36	65	9	159
Затрудняюсь ответить	1	4	2	4	1	12
Итого	41	293	215	624	42	1 215

Составлено авторами по материалам исследования

Преимуществами использования LLM для создания сводных таблиц являются высокая скорость, отсутствие ошибок, возможность построения сложных многомерных сводок, а также гибкость в формате представления результатов (таблица, текст, визуализация).

Ответ LLM (описание): «Большинство респондентов (624 человека) скорее удовлетворены своей работой, что является положительным фактором. При этом значительная доля (555 человек) среди тех, чья работа оформлена, также скорее удовлетворены. Это может свидетельствовать о том, что наличие

официального трудового договора и социальных гарантий повышает уровень удовлетворенности работой. Стоит обратить внимание на группу респондентов, чья работа не оформлена официально. Среди них наибольшее количество (65 человек) скорее удовлетворены, однако 36 человек скорее не удовлетворены, а 9 человек совсем не удовлетворены своей неоформленной работой. Это может указывать на проблемы и риски, связанные с неформальной занятостью. Интересно, что среди тех, кто затруднился ответить об оформлении работы, мнения разделились почти поровну между удовлетворенными и неудовлетворенными. В целом данные демонстрируют относительно высокий уровень удовлетворенности работой, особенно среди тех, чья занятость оформлена официально. Однако присутствует заметная группа неудовлетворенных, в том числе среди неоформленных работников».

Таким образом, LLM значительно упрощают проведение разнообразных аналитических процедур над количественными данными опросов.

СРЕДНИЕ ЗНАЧЕНИЯ

Средние значения являются одними из наиболее широко используемых статистических показателей для описания количественных данных. Традиционно расчет средних осуществляется с помощью формул или функций статистических пакетов. Однако применение LLM позволяет получать средние значения напрямую из текстовых данных опросов.

Для вычисления средних LLM достаточно указать в промте переменную, для которой требуется найти среднее, и дать инструкцию произвести расчет.

Промт может выглядеть следующим образом: «Вычисли общее среднее значение возраста респондентов по вопросу X».

LLM проанализирует текстовые ответы, извлечет числовые значения возраста и рассчитает среднее арифметическое или среднее взвешенное в зависимости от типа данных.

Ответ LLM: «В ходе опроса были собраны данные о годах рождения 1 215 респондентов. Анализ этих данных показал, что средний возраст респондентов на момент опроса составлял 46 лет. Наиболее ранний год рождения среди опрошенных – 1920, а наиболее поздний – 2001. Таким образом, в выборке присутствовали респонденты различных возрастных групп – от молодых людей до людей преклонного возраста».

Таким образом, LLM выступают универсальным инструментом для оперативного расчета и содержательного анализа средних значений по количественным данным социологических опросов.

КОРРЕЛЯЦИИ

Анализ корреляций является одним из фундаментальных методов изучения взаимосвязей между переменными в количественных исследованиях. Традиционно корреляционный анализ выполняется с использованием статистических пакетов на числовых данных. Однако с развитием технологий обработки естественного языка появилась возможность применения LLM для нахождения корреляций в текстовых данных опросов.

Для выявления корреляционных связей между переменными LLM может быть дан следующий промт: «Проанализируй данные вопросов X и Y и определи, существует ли корреляция между уровнем дохода респондентов и их удовлетворенностью жизнью. Рассчитай коэффициент корреляции и оцени его значимость».

В ответ LLM осуществит поиск упоминаний соответствующих переменных в текстовых ответах респондентов, извлечет числовые значения (если применимо) и выполнит корреляционный анализ, представив результат в виде коэффициента корреляции и оценки его статистической значимости.

Ответ LLM: «Для проведения корреляционного анализа между вопросами Q3 (удовлетворенность трудом) и Q4 (официальное оформление работы) использовался коэффициент ранговой корреляции Спирмена. Рассчитанный коэффициент корреляции Спирмена составил 0,27, что указывает на слабую положительную корреляцию между переменными».

Результаты анализа показывают, что существует слабая положительная корреляция между удовлетворенностью содержанием труда и официальным оформлением работы. Это может означать, что наличие официального трудового соглашения между работником и работодателем в некоторой степени связано с более высоким уровнем удовлетворенности работой.

Однако обнаруженная корреляция является слабой, что указывает на влияние и других факторов на удовлетворенность трудом, помимо официального оформления. Тем не менее, полученные результаты могут послужить основанием для дальнейшего изучения роли официального трудоустройства в формировании позитивного отношения работников к своей профессиональной деятельности. Необходимо также учитывать, что корреляция не означает наличие причинно-следственной связи, и для более глубокого анализа требуется привлечение дополнительных данных и теоретических концепций.

Помимо расчета коэффициентов, LLM способны генерировать описательные интерпретации обнаруженных корреляций, сравнивать их с теоретическими ожиданиями или результатами других исследований. Кроме того, они могут выявлять более сложные, нелинейные взаимосвязи между переменными.

Использование LLM для корреляционного анализа текстовых данных опросов открывает новые возможности для изучения связей между различными аспектами социальных явлений и процессов, которые ранее было сложно операционализировать в виде числовых шкал. Это расширяет границы количественных методов социологического исследования.

ЗАКЛЮЧЕНИЕ

Исследование возможностей применения больших языковых моделей для анализа количественных данных представляет собой важное направление развития методологических подходов. На основе проведенных экспериментов и теоретических изысканий можно сделать следующие выводы.

1. Универсальность и автоматизация. Модели демонстрируют высокую универсальность, обеспечивая автоматизацию анализа данных. Они эффективно обрабатывают как структурированные, так и неструктурированные данные, что особенно важно при работе с текстовыми ответами респондентов. Становится возможным оптимизация подготовки данных и снижаются риски субъективного влияния исследователя на результаты.

2. Повышение эффективности анализа. Использование моделей значительно ускоряет процесс первичного анализа данных, включая построение частотных и сводных таблиц, расчет средних значений, корреляций и другие статистические методы. Это способствует повышению точности и глубины анализа, позволяет выявлять скрытые связи и тренды, что может благоприятствовать формированию более обоснованных выводов.

3. Интерпретация и визуализация данных. Модели способны генерировать связные тексты, содержащие интерпретацию результатов анализа. С их помощью можно создавать аналитические справки, отчеты и различные графические визуализации, повышая наглядность результатов исследований.

4. Дополнение традиционных методов. Сегодня модели не заменяют, а дополняют традиционные инструменты количественного анализа данных. Совместное использование традиционных инструментов и моделей позволяет комбинировать сильные стороны обоих подходов, обеспечивая более комплексный и точный анализ данных.

5. Необходимость дальнейших исследований. Для повсеместной интеграции моделей в социологические исследования требуется провести сотни дополнительных исследований и тестов. Важно развивать методологические подходы, направленные на проверку достоверности и надежности результатов, полученных с помощью моделей.

6. Формирование инновационной культуры. Необходимо уделять внимание формированию инновационной культуры среди исследователей: повышать уровень информированности об использовании моделей, непрерывно развивать навыки работы с данными и осведомленность в технических вопросах.

Большие языковые модели – инструмент, способный значительно оптимизировать процесс анализа количественных данных в социологических исследованиях. Их применение может способствовать получению более глубоких и обоснованных выводов, что открывает новые горизонты для социологических исследований и благоприятствует развитию науки в целом.

Список литературы

1. *Brynjolfsson E., McAfee A.* The second machine age. Work, progress, and prosperity in a time of brilliant technologies. New York, London: W.W. Norton & Company; 2014. 320 p.
2. *Smith-Doerr L., Zilberstein Sh., Wilkerson T., Roberts Sh., Renski H., Green V. et al.* HTF (the future of work at the human-technology frontier). Understanding emerging technologies, racial equity, and the future of work. Alexandria: National Science Foundation; 2019. 37 p.

3. *Shestakofsky B.* Working algorithms: software automation and the future of work. *Work and Occupations*. 2017;4(44):376–423. <http://dx.doi.org/10.1177/0730888417726119>
4. *Dablin E.* Are robots stealing our jobs? *Socius*. 2019;5. <http://dx.doi.org/10.1177/2378023119846249>
5. *Felten E., Raj M., Seamans R.* Occupational, industry, and geographic exposure to artificial intelligence: a novel dataset and its potential uses. *Strategic Management Journal*. 2021;12(42):2195–2217. <http://dx.doi.org/10.1002/smj.3286>
6. *Joyce K., Smith-Doerr L., Alegria S., Bell S., Cruz T., Hoffman S.G., Noble S.U. et al.* Toward a sociology of artificial intelligence: a call for research on inequalities and structural change. *Socius*. 2021;7. <http://dx.doi.org/10.1177/2378023121999581>
7. *Brayne S.* Big data surveillance: the case of policing. *American Sociological Review*. 2017;5(82):977–1008. <http://dx.doi.org/10.1177/0003122417725865>
8. *Ламур Б.* Пересборка социального: введение в акторно-сетевую теорию. Пер. с англ. И. Полонской. М: Высшая школа экономики; 2014. 384 с.
9. *Beer D.* The social power of algorithms. *Information, Communication & Society*. 2017;1(20):1–13. <https://doi.org/10.1080/1369118X.2016.1216147>
10. *Burrell J.* How the machine “thinks”: understanding opacity in machine learning algorithms. *Big Data & Society*. 2016;1(13). <http://dx.doi.org/10.1177/2053951715622512>
11. *Pink S., Sumartojo Sh., Lupton D., La Bond H.Ch.* Mundane data: the routines, contingencies, accomplishments of digital living. *Big Data & Society*. 2017;1(4). <http://dx.doi.org/10.1177/2053951717700924>
12. *O’Neil C.* Weapons of math destruction. How big data increases inequality and threatens democracy. New York: Crown; 2016. 272 p.
13. *Enbanks V.* Automating inequality. How high-tech tools profile, police, and punish the poor. New York: St. Martin’s; 2018. 272 p.
14. *Smith-Doerr L.* Hidden injustice and anti-science. *Engaging Science, Technology, and Society*. 2020;6:94–101. <http://dx.doi.org/10.17351/ests2020.381>
15. *Vallas S., Schor J.B.* What do platforms do? Understanding the gig economy. *Annual Review of Sociology*. 2020;46:273–294. <http://dx.doi.org/10.1146/annurev-soc-121919-054857>
16. *Шавель С.А.* Социальные инновации в жизни общества. *Наука и инновации*. 2007;4:10–12.
17. *Круглов В.В., Али М.И.* Применение аппарата нейронных сетей для анализа социологических данных. *Социологические исследования*. 2001;9:112–114.
18. *Шакирова А.Ф.* Особенности применения нейросетевого и пространственного анализа в социологических науках на примере построения индекса социальных настроений жителей города Казани. В кн.: История, политология, социология, философия: теоретические и практические аспекты: сборник статей по материалам XXIII международной научно-практической конференции, том 8(16), Новосибирск, 5 августа 2019 г. Новосибирск: Ассоциация научных сотрудников «Сибирская академическая книга»; 2019. С. 47–50.
19. *Фомина Е.Е.* Применение аппарата нейронных сетей для анализа результатов анкетирования. *Вестник Пермского национального исследовательского политехнического университета. Социально-экономические науки*. 2020;2:99–110. <http://doi.org/10.15593/2224-9354/2020.2.8>
20. *Мальцева А.В., Шилкина Н.Е., Махныткина О.В.* Data mining в социологии: опыт и перспективы проведения исследования. *Социологические исследования*. 2016;3:35–44.

References

1. *Brynjolfsson E., McAfee A.* The second machine age. Work, progress, and prosperity in a time of brilliant technologies. New York, London: W.W. Norton & Company; 2014. 320 p.
2. *Smith-Doerr L., Zilberstein Sh., Wilkerson T., Roberts Sh., Renski H., Green V. et al.* HTF (the future of work at the human-technology frontier). Understanding emerging technologies, racial equity, and the future of work. Alexandria: National Science Foundation; 2019. 37 p.
3. *Shestakofsky B.* Working algorithms: software automation and the future of work. *Work and Occupations*. 2017;4(44):376–423. <http://dx.doi.org/10.1177/0730888417726119>
4. *Dablin E.* Are robots stealing our jobs? *Socius*. 2019;5. <http://dx.doi.org/10.1177/2378023119846249>
5. *Felten E., Raj M., Seamans R.* Occupational, industry, and geographic exposure to artificial intelligence: a novel dataset and its potential uses. *Strategic Management Journal*. 2021;12(42):2195–2217. <http://dx.doi.org/10.1002/smj.3286>
6. *Joyce K., Smith-Doerr L., Alegria S., Bell S., Cruz T., Hoffman S.G., Noble S.U. et al.* Toward a sociology of artificial intelligence: a call for research on inequalities and structural change. *Socius*. 2021;7. <http://dx.doi.org/10.1177/2378023121999581>
7. *Brayne S.* Big data surveillance: the case of policing. *American Sociological Review*. 2017;5(82):977–1008. <http://dx.doi.org/10.1177/0003122417725865>

8. *Latour B.* Reassembling the social: an introduction to actor-network-theory. Trans. from Eng. I. Polonskaya. Moscow: Higher School of Economics; 2014. 384 p. (In Russian).
9. *Beer D.* The social power of algorithms. *Information, Communication & Society.* 2017;1(20):1–13. <https://doi.org/10.1080/1369118X.2016.1216147>
10. *Burrell J.* How the machine “thinks”: understanding opacity in machine learning algorithms. *Big Data & Society.* 2016;1(13). <http://dx.doi.org/10.1177/2053951715622512>
11. *Pink S., Sumartojo Sb., Lupton D., La Bond H.Ch.* Mundane data: the routines, contingencies, accomplishments of digital living. *Big Data & Society.* 2017;1(4). <http://dx.doi.org/10.1177/2053951717700924>
12. *O’Neil C.* Weapons of math destruction. How big data increases inequality and threatens democracy. New York: Crown; 2016. 272 p.
13. *Eubanks V.* Automating inequality. How high-tech tools profile, police, and punish the poor. New York: St. Martin’s; 2018. 272 p.
14. *Smith-Doerr L.* Hidden injustice and anti-science. *Engaging Science, Technology, and Society.* 2020;6:94–101. <http://dx.doi.org/10.17351/ests2020.381>
15. *Vallas S., Schor J.B.* What do platforms do? Understanding the gig economy. *Annual Review of Sociology.* 2020;46:273–294. <http://dx.doi.org/10.1146/annurev-soc-121919-054857>
16. *Shavel S.A.* Social innovations in the life of society. *Science and Innovation.* 2007;4:10–12. (In Russian).
17. *Kruglov V.V., Dli M.I.* Application of neural network tools for the analysis of sociological data. *Sociological Studies.* 2001;9:112–114. (In Russian).
18. *Shakirova A.F.* Features of the use of neural network and spatial analysis in the sociological sciences on the example of building an index of social moods of residents of the city of Kazan. In: *History, political science, sociology, philosophy: theoretical and practical aspects: Proceedings of the XXIII International Scientific and Practical Conference, volume 8(16), Novosibirsk, August 5, 2019.* Novosibirsk: Association of Researchers “Siberian Academic Book”; 2019. Pp. 47–50. (In Russian).
19. *Fomina E.E.* Application of neural network tools for analysis of the survey results. *PNRPU Sociology and Economics Bulletin.* 2020;2:99–110. (In Russian). <http://doi.org/10.15593/2224-9354/2020.2.8>
20. *Maltseva A.V., Shilkina N.E., Makhnutkina O.V.* Data mining in sociology: experience and outlook for research. *Sociological Studies.* 2016;3:35–44. (In Russian).