Мария Михайловна Базлуцкая

Московский государственный университет международных отношений Министерства иностранных дел Российской Федерации

m.bazlutskaya@gmail.com

Анна Николаевна Сытник

Санкт-Петербургский государственный университет

a.sytnik@spbu.ru

**Рамочный анализ данных на русском языке с помощью ChatGPT**

На примере анализа коротких текстовых сообщений авторы демонстрируют возможности применения генеративного искусственного интеллекта ChatGPT к индивидуальному рамочному анализу. Статья призвана найти пути, которые могли бы привести к избавлению ученых от трудностей, связанных с ручным кодированием лексических конструкций и автоматическим кодированием фреймов на русском языке. Авторы предлагают заменить поиск фреймов и тематическое моделирование при помощи языков Python или R, требующих соответствующих навыков и не адаптированных к русскому языку, на аналогичное взаимодействие с искусственным интеллектом с целью повышения качества исследования, облегчение процесса и снижения «человеческого фактора».

Ключевые слова: искусственный интеллект, рамочный анализ, тематическое моделирование, фрейм.

В 2022 г. произошла революция генеративного искусственного интеллекта (ИИ). Выпуск ChatGPT-3.5 привлек научное сообщество возможностями на стыке «сотрудничества» ученого и ИИ. Назовем лишь две причины, почему подобные языковые модели интересны для расширения исследовательского потенциала. Первая – снижение порога входа в те научные направления, где развиваются методики анализа данных. Вторая – взаимодействие со сложнейшими языковыми моделями, обученными на данных «всего мира». Доступ к диалоговому окну позволяет ученому «прикоснуться» к ним, интерпретировать собранные данные без вовлечения в технологический процесс и поэкспериментировать с предвзятостью самих моделей. В статье мы продемонстрируем это на примере анализа твитов с политическим контекстом. Мы задаем вопрос, насколько успешно ChatGPT 3.5/4 версий справится с выявлением тем и рамочным анализом твитов на русском языке?

Классическое определение фреймов в 1993 г. дал Р. Энтман [2]. Его работу дополняют исследования Гэмсона и Модильяни [3], Али и Хассана [1] и др. В нашей работе мы остановимся на поиске фреймов и тематическом моделировании (ТМ) текста.

Самая ресурсозатратная составляющая – выявление рамочных конструкций. Привычный метод – ручное кодирование при условии высокой степени согласия кодировщиков. Несмотря на то, что он проверен и опробован в ряде работ [4], сохраняется вероятность предвзятого анализа. Альтернатива – автоматизированный машинный анализ. Последний вместе с ТМ могут быть выполнены на языках Python/R. Созданные для них библиотеки заявляют о способности к поиску тем и определению особенных лексических конструкций в тексте. Однако для исследователей, не владеющих навыками программирования, это преграда.

Попыткой избежать предвзятости и технологическими препятствиями и было вызвано наше обращение к инструменту от разработчиков OpenAI – ChatGPT. Мы предположили, что, поскольку Chat распознает русский язык, натренирован в технике обучения с подкреплением на основе отзывов людей с использованием супервычислительной инфраструктуры Microsoft Azure, то, вероятно, сможет помочь определить фреймы и распознает темы твитов.

Мы задали ИИ вопрос, может ли он это сделать, получили положительный ответ и провели тестовый анализ. По его итогам программа определила речевые конструкции, которые, по определению Энтмана, можно считать фреймами, объяснила свой выбор и выделила темы каждого сообщения (это важный момент, так как при тематическом моделировании с Python/R проблема возникает при интерпретации результатов). Итоги анализа были представлены в таблице.

Задание вводилось на английском языке, чтобы программа лучше его поняла. Результат выводился на русском языке для удобства исследователей. Для отработки данного метода мы использовали текстовые сообщения из социальной сети Twitter\*[[1]](#footnote-1), их объем составил 3000 уникальных твитов.

Мы смогли обработать в программе 99,85% введенных данных. В остальных 0,15% (45 сообщениях) Chat не смог выполнить задание полностью из-за недостаточности материала для определения темы или отсутствия рамочных конструкций в тексте.

Несмотря на то, что программа выполнила поставленную перед ней задачу, исследователи столкнулись с рядом ограничений:

1. Chat GPT версий 3.5/4 не осуществляет одномоментный анализ данных, объем которых превышает 4096 токенов. Это вынуждает исследователей вводить массив порционно и снижает скорость исследования.
2. Выводимые данные требуют контроля со стороны кодировщика, т.к. Chat может проигнорировать запрос или пропустить фреймы. Для устранения проблемы кодировщик должен задавать программе дополнительные вопросы.
3. Фреймы на русском языке выводятся, как указано в исходном тексте, даже в случае требования привести фразы к инфинитиву. Это вынуждает обрабатывать результаты вручную.

Предлагаемый способ рамочного анализа и ТМ проще языков Python/R в использовании, позволяет снизить предвзятость исследователя и может быть в применен к другим видам анализа, которые ранее были недоступны для большинства ученых в гуманитарных науках из-за технической сложности.

Литература

1. Ali M., Hassan N. A Survey of Computational Framing Analysis Approaches // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022. P. 9335–9348.

2. Entman R. M. Framing: Toward clarification of a fractured paradigm // Journal of communication. 1993. Vol. 43. No 4. P. 51–58.

3. Gamson W. A., Modigliani A. Media discourse and public opinion on nuclear power: A constructionist approach // American journal of sociology. 1989. Vol. 95. No 1. P. 1–37.

4. Semetko H. A., Valkenburg P. M. Framing European politics: A content analysis of press and television news // Journal of communication. 2000. Vol. 50. No 2. P. 93–109.

1. Организация признана экстремистской по решению суда, деятельность организации запрещена на территории РФ. [↑](#footnote-ref-1)