**«Есть ли ценности в посте?» – Извлечение ценностно-ориентированных текстов из социальных медиа**

*Милкова М.А., Руднев М.Г., Окольская Л.А.*

Изучение жизненных ценностей на основе данных цифровых следов в социальных сетях является актуальной, но крайне сложной для решения задачей в области вычислительных социальных наук. В предшествующих работах основной интерес исследователей в данной области лежал в сопоставлении слов, используемых людьми и результатами стандартного для психологии измерения ценностей с помощью формализованного опросника (Park et al., 2015; Kulkarni et al., 2018). Многочисленные попытки поиска корреляций между ценностными ориентациями и их выражением в языке демонстрируют лишь слабые связи (Ponizovskiy et al., 2020; Fischer et al., 2022). Причина этого несоответствия заключается в том, что выражение ценностей в текстах является поведенческим индикатором ценностных ориентаций, который качественно отличается от ответов на вопросы анкет. В частности, пользовательский контент в социальных сетях в большей степени связан с массовым использованием шаблонных, культурно заданных речевых конструкций, чем стремлением выразить собственные ценности. Таким образом, актуальной и ранее не решенной задачей является выделение ценностно-выражающего (value-expressive) контента, присутствующего в социальных сетях и последующий анализ его *восприятия*.

Наша работа ставит перед собой цель автоматического определения текстов, имеющих ценностную ориентацию. Мы фокусируемся на российской сети ВКонтакте и стремимся представить репрезентативную выборку пользователей. На основе постов случайно отобранных пользователей мы строили модели машинного обучения, в том числе, глубокого обучения, для классификации текстов на предмет наличия в них ценностных ориентаций.

Пилотное исследование включало посты 5621 пользователей, общим объемом 188246 текстовых записей. Анализ постов пользователей, которые отобраны случайно, без каких-либо предустановленных характеристик, имеет свою специфику. Так, ключевой особенностью постов в социальной сети ВКонтакте является сильная зашумленность данных, выражающаяся в большом числе спам-сообщений: разнообразных рекламных текстов, а также сообщений, которые генерируются автоматически после использования пользователями ВК-приложений. Сложность выявления спам-сообщений заключается, во-первых, в их разнообразии (широкий спектр рекламных сообщений в различных отраслях, включая нативную рекламу), во-вторых, в «натуралистичных» формулировках автоматически сгенерированных приложениями сообщений.

На первом этапе мы проводили фильтрацию данных от спама, в несколько итераций. Сначала с использованием тематического моделирования были выделены ключевые фразы, характеризующие спам-сообщения. На основе выделенных фраз, а также некоторых других правил была получена нечеткая разметка данных, которая использовалась для последующей настройки моделей машинного обучения, на основе эмбеддингов слов. Полученная классификация (F1=0.92) позволила выделить посты, не содержащие спама (146304 записей или 78% от исходной выборки) и представляющие интерес для дальнейшего анализа.

На втором этапе мы проводим бинарную классификацию и не различаем посты по конкретным видам ценностей. Для разметки и настройки модели глубокого обучения (русскоязычной разговорной дистиллированной модели BERT – Kolesnikova et al., 2022), а также тестирования прочих моделей машинного обучения с целью классификации постов мы использовали подход Active Learning. Это подход к разметке данных, когда разметка осуществляется не случайным образом, а представляет из себя итеративный процесс. На стартовом этапе специально обученный человек размечает небольшое число данных. Затем запускается модель обучения, которая осуществляет классификацию. По итогам ее работы человеку остается разметить только те данные, в которых модель не уверена (вероятности классов около 0.5). Процесс повторяется до достижения приемлемого качества классификации.

Полученный массив данных ценностно-ориентированных постов позволяет, во-первых, количественно оценить их представленность в сети ВКонтакте, во-вторых, открывает возможности для дальнейших исследований по анализу восприятия ценностных ориентаций, не прибегая к трудоемкому ручному сбору данных.

**Литература**

Fischer, R., Karl, J., Fetvadjiev, V., Grener, A., Luczak-Roesch M. (2022). Opportunities and Challenges of Extracting Values in Autobiographical Narratives // Frontiers in Psychology, 13:886455. <https://doi.org/10.3389/fpsyg.2022.886455>

Kolesnikova, A., Kuratov, Y., Konovalov, V., Burtsev, M. (2022). Knowledge Distillation of Russian Language Models with Reduction of Vocabulary // arXiv:2205.02340 <https://doi.org/10.48550/arXiv.2205.02340>

Kulkarni, V., Kern, M.L. Stillwell, D., Kosinski, M., Matz, S., Ungar, L., Skiena, S., Schwartz, H.A. (2018). Latent human traits in the language of social media: An open-vocabulary approach // PLoS ONE 13(11): e0201703. <https://doi.org/10.1371/journal.pone.0201703>

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., Seligman, M. E. P. (2015). Automatic personality assessment through social media language // Journal of Personality and Social Psychology, 108(6), pp. 934–952. <https://doi.org/10.1037/pspp0000020>

Ponizovskiy, V., Ardag, M., Grigoryan, L., Boyd, R., Dobewall, H., Holtz, P. (2020). Development and Validation of the Personal Values Dictionary: A Theory-Driven Tool for Investigating References to Basic Human Values in Text // European Journal of Personality, 34(5), pp. 885–902. <https://doi.org/10.1002/per.2294>