Идентификация человеческих ценностей в текстовых аргументах

М. В. Лаптев, email: mv@paq.su
С. В. Вычегжанин, email: vychegzhaninsv@gmail.com
Е. В. Котельников, email: kotelnikov.ev@gmail.com

ФГБОУ ВО «Вятский государственный университет»

Аннотация. *В работе рассматриваются человеческие ценности, такие как «Самостоятельность», «Забота о природе» и т.д., и осуществляется попытка идентификации этих ценностей в текстовых аргументах на русском языке. Предоставляется перевод на русский язык размеченного англоязычного текстового корпуса с разметкой аргументов по ценностям. Достигнуто качество $F_1 = 0{,}4034$, что является улучшением относительно текущих результатов в англоязычном сегменте.*

Ключевые слова: *человеческие ценности, текстовые аргументы, идентификация, TF-IDF, Word2Vec, BERT.*

Введение

Почему люди расходятся во мнениях относительно наиболее правильного решения спорных вопросов даже при условии использования одной и той же информации для формирования собственного мнения? Для ответа на этот вопрос необходимо узнать факторы, формирующие мнение отдельного человека. Известно, что разные люди имеют разные мнения и приоритеты в отношении того, чего следует достигать (стремление быть успешным / быть скромным) и что делать (придерживаться традиций / быть самостоятельным), что в совокупности называют человеческими ценностями [1]. Согласно определению, человеческие ценности – это связанные с желаемым конечным состоянием или способом поведения убеждения, выходящие за рамки конкретных ситуаций, оценивающие поведение людей и события, а также упорядоченные по важности друг относительно друга, формируя систему ценностных приоритетов [2]. Таким образом, забота о природе является ценностью, а предпочтения в еде – нет.

Задача идентификации ценностей в текстовых аргументах является сложной из-за большого количества ценностей, неявных отсылок к ним в аргументах и неоднозначности аргументов в естественном языке. Однако создание текстовых корпусов с разметкой по аргументации и

прогресс последних лет в обработке естественного языка позволяют решать подобного рода задачи.

В настоящей статье впервые для русского языка выполняется исследование качества автоматической идентификации ценностей в аргументационных текстах на основе методов машинного обучения. Исследование проводится на материале англоязычного текстового корпуса с разметкой 5220 аргументов по ценностям, переведенного на русский язык. Данный корпус предоставлен в общий доступ.

1. Использование ценностей для анализа аргументов

Большинство социальных наук так или иначе взаимодействуют с человеческими ценностями. М. Рокич утверждает, что ценности являются убеждениями, относящимися к конечным состояниям или способам поведения, а системы ценностей – это расстановки приоритетов ценностей, основанные на личных, культурных и социальных факторах, приписывая ценности людям, а не объектам [3]. Данная статья следует этим определениям, делая акцент на личные ценности, стоящие за аргументами.

Связь между человеческими ценностями и аргументами находила применение и ранее. Например, человеческие ценности использовались в вычислительных системах аргументации [4] или анализировались для профилирования личности, используя пользовательские эссе и тексты в социальных сетях [5].

В настоящее время широкое распространение получила теория базовых индивидуальных ценностей Ш. Шварца, схема которой представлена на рис. 1 [6]. Ш. Шварц основывал порядок расположения ценностей на базе конфликта или совместимости различных ценностей, которые совместно проявляются в определенных ситуациях. Например, решение бросить вызов власти создает конфликт между «Самостоятельностью» и «Конформизмом», но способствует проявлению и «Самостоятельности», и «Стимуляции» [6]. Из этого можно сделать вывод, что человеческие ценности могут обеспечить контекст для категоризации, сравнения и оценки аргументов, позволяя обеспечивать исследования в социальных науках информацией о ценностях с помощью крупномасштабных наборов данных; оценивать аргументацию по масштабу и силе; генерировать или подбирать аргументы, исходя из системы ценностей целевой аудитории; выявлять противоположные и общие ценности с обеих сторон спорной темы [7].

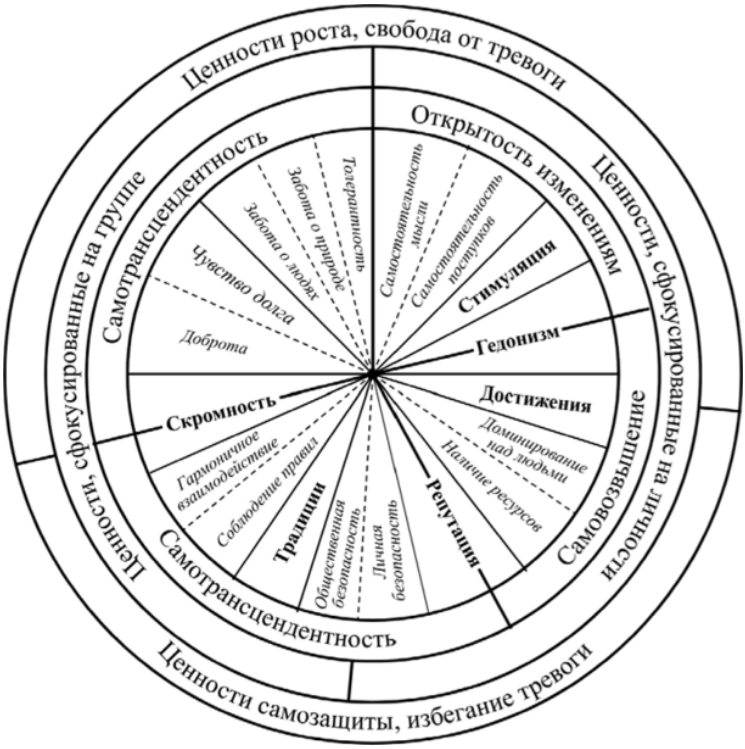

Рис. 1. Схема ценностей Ш. Шварца [8]

В табл. 1 приведены 19 базовых ценностей Шварца с потенциально различным мотивационным смыслом и дается определение каждой из них в терминах мотивационных целей [6].

Таблица 1

Ценности в уточненной теории Ш. Шварца

Ценность	**Концептуальное определение с точки зрения мотивационной цели**
Самостоятельность – Мысли	Свобода развивать собственные идеи и способности
Самостоятельность – Поступки	Свобода определять собственные действия
Стимуляция	Стремление к возбуждению, новизне и переменам

Окончание табл. 1

Ценность	**Концептуальное определение с точки зрения мотивационной цели**
Гедонизм	Стремление к удовольствию и чувственному удовлетворению
Достижение	Достижение успеха в соответствии с социальными стандартами (нормами)
Власть – Доминирование	Влияние посредством осуществления контроля над людьми
Власть – Ресурсы	Влияние посредством контролирования материальных и социальных ресурсов
Репутация	Защита и влияние посредством поддержания публичного имиджа и избегания унижения
Безопасность – Личная	Безопасность непосредственного окружения
Безопасность – Общественная	Безопасность и стабильность общества в целом
Традиция	Поддержание и сохранение культурных, семейных или религиозных традиций
Конформизм – Правила	Соблюдение правил, законов и формальных обязательств
Конформизм – Межличностный	Избегание причинения вреда или огорчения другим людям
Скромность	Признание незначительности существования одного человека в круговороте жизни
Универсализм – Забота о других	Стремление к равенству, справедливости и защите всех людей
Универсализм – Забота о природе	Сохранение природной среды
Универсализм – Толерантность	Принятие и понимание тех, кто отличается от тебя
Благожелательность – Забота	Преданность группе и благополучие ее членов
Благожелательность – Чувство долга	Стремление быть надежным и заслуживающим доверия членом группы

2. Текстовый корпус

В этом разделе представлен текстовый корпус для изучения человеческих ценностей, стоящих за аргументами[1]. Корпус представляет собой перевод на русский язык единственного существующего англоязычного набора данных из 5220 аргументов и сопоставленных им ценностей [7]. Каждый аргумент состоит из одного утверждения, одного вывода и позиции аргумента по отношению к утверждению («за» или «против»). В табл. 2 приведены примеры аргументов с сопоставленными этим аргументам ценностями.

Таблица 2

Утверждение	**Вывод**	**Позиция**	**Ценности**
Школьная форма лишает детей возможности самовыражения через выбор одежды.	Мы должны отказаться от использования школьной формы.	за	Самостоятельность – Мысли Самостоятельность – Поступки
Наличие ядерного оружия может привести к определенному уровню уважения.	Мы должны бороться за отмену ядерного оружия.	против	Репутация
Либертарианство способствует более справедливому и эффективному обществу для всех.	Мы должны принять либертарианство.	за	Безопасность – Общественная Универсализм – Забота о других

Важным дополнением является тот факт, что создатели корпуса добавили дополнительную категорию к используемым категориям Шварца – «Универсализм: объективность», означающую наличие объективного взгляда и следование логическим рассуждениям. На рис. 2 приведена диаграмма частоты встречаемости ценностей среди аргументов корпуса. Наиболее распространенными ценностями оказались «Забота о других», «Личная безопасность» и «Общественная безопасность». Количество символов в утверждениях в корпусе варьируется от 24 до 766, составляя в среднем 124 символа.

[1] https://github.com/Blastieq/humanValueDetection.

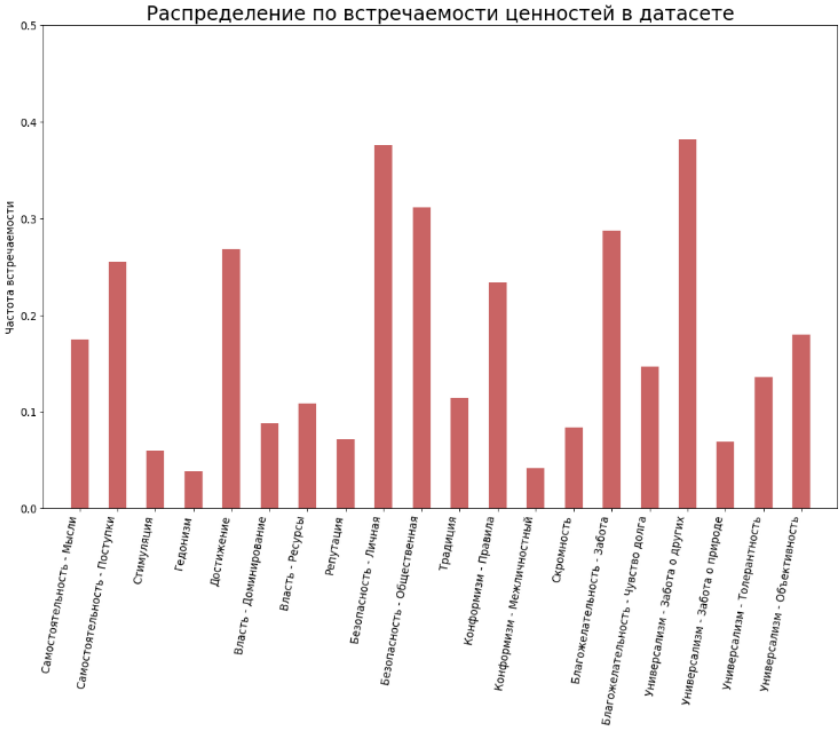

Рис. 2. Распределение человеческих ценностей в корпусе по встречаемости

3. Идентификация ценностей, стоящих за аргументами

В этом разделе представлена попытка автоматической идентификации человеческих ценностей в аргументах для русского языка с использованием машинного обучения. Сравниваются следующие подходы:

– подход с формированием TF-IDF матрицы и дальнейшей классификацией при помощи следующих методов машинного обучения: логистической регрессии, метода опорных векторов, метода k ближайших соседей, дерева решений, случайного леса, градиентного бустинга;

– подход с использованием предобученной Word2Vec модели из библиотеки Navec [9] и дальнейшей классификацией при помощи указанных методов машинного обучения;

– подход с использованием тонкой настройки предобученных русскоязычных моделей семейства BERT.

Для первых двух подходов исследуются отдельно варианты с лемматизацией текстов на основе библиотеки pymorphy2 [10] и без лемматизации.

Модели оцениваются по F_1-мере, считая среднее значение по всем классам (macro-averaging). Мера выбрана таким образом, чтобы придать всем ценностям одинаковый вес.

В табл. 3 приведены лучшие результаты для сравниваемых подходов. Максимальный результат показал подход с использованием тонкой настройки предобученной русскоязычной модели RuBERT на основе исходного текста. Примечательно, что по сравнению с предыдущей работой англоязычных коллег [7] (F_1=0.34) результат классификации улучшился, однако это может быть связано как с особенностями русского языка, так и с увеличением количества рассматриваемых подходов

Таблица 3

Результаты классификации аргументов по ценностям

Используемый подход	**F_1-мера**
Дерево решений для TF-IDF матрицы на основе лемматизированного текста	0.3887
Метод k ближайших соседей для Word2Vec эмбеддингов на основе лемматизированного текста	0.3885
Дообученная за 12 эпох модель Deeppavlov/rubert-base-cased на основе исходного текста	0.4034

Также стоит упомянуть тот факт, что за исключением нескольких ценностей прослеживается определенная корреляция между частотой встречаемости ценности в корпусе и качеством идентификации этой ценности, что показано на рис. 3. Это говорит о том, что размер корпуса на текущий момент недостаточно велик, поэтому в первую очередь для дальнейших исследований в этом направлении необходимо решить вопрос с увеличением набора данных.

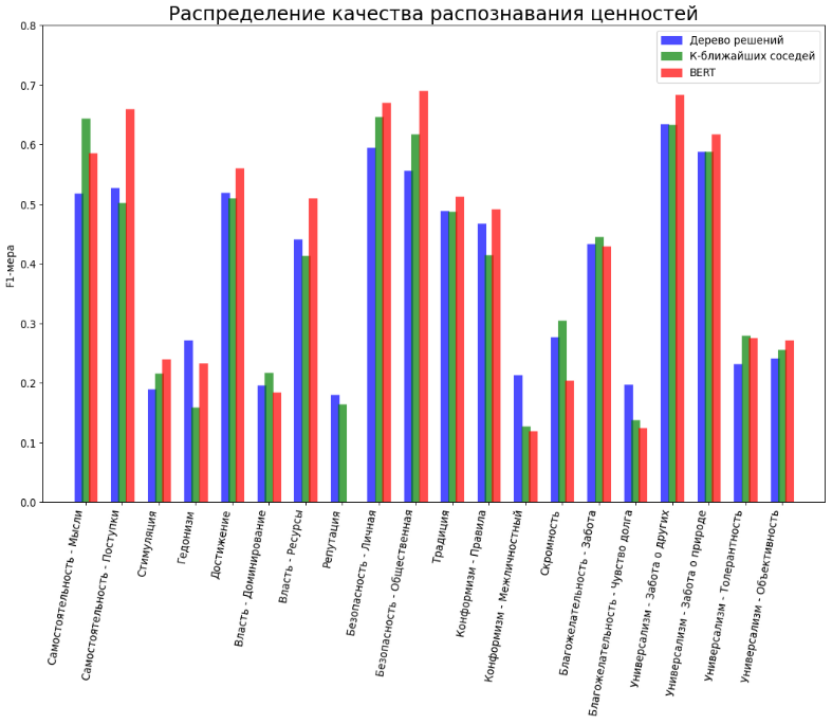

Рис. 3. Результат качества распознавания ценностей на основе лучших из рассмотренных подходов

Заключение

Задача идентификации человеческих ценностей по текстовым аргументам является сложной, однако решение этой задачи открывает большое количество возможностей, в том числе, позволяет обеспечивать исследования в социальных науках информацией о ценностях с помощью крупномасштабных наборов данных, оценивать аргументацию по масштабу и силе, генерировать или подбирать аргументы, исходя из системы ценностей целевой аудитории, выявлять противоположные и общие ценности с обеих сторон спорной темы.

Данная статья является первой работой, посвященной проблеме идентификации человеческих ценностей в русскоязычных текстовых аргументах. Предоставлен переведенный с английского языка текстовый корпус, размеченный по ценностям, и сделана попытка автоматической идентификации человеческих ценностей для русского языка при

помощи машинного обучения. Рассмотренные подходы показали более высокие результаты, чем в предыдущих работах англоязычных коллег [7], что может быть связано как с методологией, так и с особенностями русского языка.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 22-21-00885, https://rscf.ru/project/22-21-00885/.

Список литературы

1. Searle, J. R. Rationality in Action / MIT Press, 2003.
2. Schwartz, S. H. Are There Universal Aspects in the Structure and Contents of Human Values? / S. Schwartz // Journal of Social Issues. – 1994. – Vol. 50. – P. 19–45.
3. Rokeach, M. The Nature of Human Values / New York: The Free Press, 1973.
4. Bench-Capon, T. J. M. Persuasion in Practical Argument Using Value-based Argumentation Frameworks / T. J. M. Bench-Capon // Journal of Logic and Computation – 2003. – Vol. 13(3). – P. 429–448.
5. Maheshwari, T. A Societal Sentiment Analysis: Predicting the Values and Ethics of Individuals by Analysing Social Media Content / T. Maheshwari, A. N. Reganti, S. Gupta et al. // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017). – Valencia, 2017. – P. 731–741.
6. Schwartz, S. H. Refining the theory of basic individual values / S. H.Schwartz, J. Cieciuch, M. Vecchione et al. // Journal of personality and social psychology. – 2012. – Vol. 4. – P. 663–688.
7. Kiesel, J. Identifying the Human Values behind Arguments / J. Kiesel, M. Alshomary, N. Handke et al. // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics – Dublin, 2022. – P. 4459–4471.
8. Шварц, Ш. Уточненная теория базовых индивидуальных ценностей: применение в России / Ш. Шварц, Т. П. Бутенко // Психология. Журнал Высшей школы экономики. – 2012. – № 1. – С 43–70.
9. Navec: компактные эмбеддинги для русского языка [Электронный ресурс] : база данных. – Режим доступа : https://natasha.github.io/navec.
10. Korobov, M. Morphological Analyzer and Generator for Russian and Ukrainian Languages / M. Korobov // Proceedings of the 4th International Conference “Analysis of Images, Social Networks and Texts” (AIST 2015). – Yekaterinburg, 2015. – P. 320–332.