РАЗМЕТКА ТЕКСТОВЫХ МАССИВОВ

TEXTS MARKUP

УДК 81’33 DOI: 10.31249/chel/2025.02.03

Колмогорова А.В., **Хлебникова В.А.**

ПОМОЖЕТ ЛИ БАЙЕСОВСКАЯ СЫВОРОТКА ПРАВДЫ ПОВЫСИТЬ ДОСТОВЕРНОСТЬ РАЗМЕТКИ ЭМОЦИОНАЛЬНЫХ ТЕКСТОВ? (CASE STUDY)[©, 1]

Национальный исследовательский университет «Высшая школа экономики», ООО «Яндекс Крауд», Россия, Санкт-Петербург, akolmogorova@hse.ru, va.khleb@yandex.ru

Аннотация. В статье рассматриваются результаты применения методологии, известной как Байесовская сыворотка правды (BTS), в эмоциональной разметке текстов для последующего обучения нейросетевых моделей. Суть метода состоит в том, что информантов сначала просят оценить некоторый феномен со своей собственной точки зрения, а затем – предсказать, какой ответ (или оценку) выберет наибольший процент других отвечающих на тот же опросник. Мы применили данную методологию для оценки 120 разметчиками 300 эмоциональных текстов, извлеченных из группы «Подслушано» социальной интернет-сети ВКонтакте, где они имели эмоциональные хештеги. В основе дизайна разметки лежала PAD-модель Рассела – Мехрабиана. При обработке результатов сравнивались средние значения стандартного отклонения в личных и предсказанных оценках по каждой из трех шкал модели. Сформировав подкорпусы текстов с наибольшей рассогласованностью личной и предсказанной оценок, мы проанализировали их,

[1] Статья подготовлена по материалам проекта «Текст как Big Data: методы и модели работы с большими текстовыми данными», выполняемого в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 г.

выявив частотные слова для каждого из подкорпусов. Получены следующие выводы: 1) разброс личных оценок и предсказанных оценок в собранном датасете не имеет статистически значимых отличий; 2) в подкорпусы текстов с наибольшим расхождением личной и предсказанной эмоциональной оценки попадают тексты, посвященные трем типам социальных ситуаций: взаимоотношения внутри пары, отношения мать – ребенок, а также девиантное поведение, подвергающее риску безопасность семьи и других членов социума; 3) наибольшее число текстов, в которых наблюдается значимое расхождение оценок, маркированы хештегами, связанными с эмоциями страха, отвращения, удивления, воодушевления и грусти.

Ключевые слова: эмоциональные тексты; разметка; Байесовская сыворотка правды; детектирование эмоций; модель эмоций PAD.

Получена: 10.09.2024 Принята к печати: 28.12.2024

Kolmogorova A.V., Khlebnikova V.A.

Will Bayesian truth serum help to increase the reliability of markup of emotional texts? (case study)[©, 1]

National Research University "Higher School of Economics", Yandex Crowd LLC, Russia, Saint-Petersburg, akolmogorova@hse.ru, va.khleb@yandex.ru

Abstract. The paper examines the results of a methodology known as Bayesian Truth Serum (BTS) when applied in the task of emotional markup of texts for neural network models training. Bayesian truth serum is used in experimental surveys for which the category of truth is not applicable – the results obtained in such surveys cannot be verified by comparing them with a certain standard, since the latter does not exist. The framework presupposes that informants are first asked to evaluate a certain phenomenon from their own point of view, and then – to predict which answer (or assessment) the largest percentage of other respondents to the same questionnaire will choose. Emotional markup of texts is also a task in which we do not know the true emotion, but we can stimulate the truthfulness of informants using the BTS method. We applied this methodology to evaluate 300 emotional texts retrieved from the group "Overheard" on VKontakte. 120 informants took part in the emotional markup procedure carried out with tenfold coverage. The markup design was based on the Russell–Mehrabian PAD model: informants were asked to assess what emotions the author of the text was experiencing, using three eleven–point scales (Pleasure, Arousal, Dominance). When processing the results, we compared the average values of the standard deviation in personal and predicted estimates given by the informants on each of the

[1] The article was prepared based on the materials of the project "Text as Big Data: Methods and Models of Working with Big Text Data", which is carried out within the framework of the Fundamental Research Program of the National Research University Higher School of Economics (HSE University) in 2024.

three scales. Then we formed a subcorpus of texts with the greatest inconsistency of personal and predicted estimates and analyzed the words frequency in each of the subcorpora. In addition, according to the emotional tags under which the texts were published in VK, we calculated the "weight" of eight emotions for these texts with the greatest discrepancies. The main hypothesis was that the greatest inconsistency of personal and predicted assessment will be found in those texts that describe a certain situation of social interaction, for which the demonstration of emotional behavior is regulated by some implicit rules. The study resulted in the following conclusions: 1) the spread in the set of personal estimates and predicted ones has no statistically significant differences; 2) texts having the greatest discrepancy between personal and predicted emotional assessment concern three main type of situations: relationships within a couple (husband – wife, boyfriend – girlfriend), mother-child relationships and deviant behavior that poses a threat to the safety of the family and other members of society; 3) the largest number of texts showing most of the discrepancies in ratings, is marked with emotional hashtags associated with the emotions of fear, disgust, surprise, excitement and sadness. The prospect of the study is the continuation of experiments with markup on different samples of informants.

Keywords: emotional texts; markup; Bayesian truth serum; emotion detection; emotion PAD model.

Received: 10.09.2024 Accepted: 28.12.2024

Введение

С появлением ChatGPT не только для специалистов, но и для широкого круга людей стала ясна основная идея разработчиков «сильного» ИИ – сделать его максимально антропоморфным. При этом границы достижимого уровня антропоморфизма существенно расширились – сегодня это не только понимание естественного языка (*Natural Language Understanding*), извлечение информации (*Information Extraction*), но и задачи, связанные с автоматической интерпретацией поведенческих паттернов и детектированием скрытых психологических состояний у человека (*Affective Computing*). Тем не менее, несмотря на общую амбициозность таких целей, отправной точкой для их достижения по-прежнему пока является размеченный датасет, на котором нейросеть или LLM (*Large Language Model*) дообучается. В таком датасете некоторое множество объектов представлено в формате «объект – ключ». Однако при решении задач высокой степени антропоморфизма (то есть требующих привлечения специфически человеческого когнитивного опыта, получить который можно только на основании жизненного

опыта) для того, чтобы иметь «ключи», то есть некоторые признаки интересующих нас объектов, мы нуждаемся в размеченных людьми данных и приглашаем для этого информантов. Важная проблема, которая возникает при разметке для задач высокой степени антропоморфизма, состоит в том, что мы не можем оценить ее качество, потому что не знаем правильных ответов – у нас нет так называемой *ground trough*.

Сравним две задачи разметки: 1) разметить, какой объект изображен на картинке (например, для обучения модели компьютерного зрения) и 2) разметить, какую эмоцию хотел выразить автор текста (для обучения модели эмоциональному анализу текстов). Если в первом случае правильный ответ понятен на уровне здравого смысла и экспериментатору, и разметчику, если у них нет неврологических патологий или патологий зрения, то во втором случае у нас нет правильного ответа, поскольку слишком много факторов – языковых, психологических, когнитивных и т.д. – предопределяют оценку эмоции в каждом конкретном случае. При этом один из самых сильных предикторов, влияющих на оценку наблюдаемой эмоции, – это социальная предпочтительность эмоции в той или иной ситуации. Как отмечает П. Шародо [Charaudeau, 2000], «правильная» с точки зрения группы, в которую входит субъект, эмоциональная реакция – подтверждение субъектом своей принадлежности к группе и ее ценностям. Иными словами, недостаточно просто испытывать эмоцию, чтобы ее продемонстрировать, манифестировать, субъект должен быть уверен, что с точки зрения его микросоциума она легитимна. Так, демонстрировать свой гнев по отношению к ребенку в сообществах современных «осознанных» родителей нелегитимно, а в советских семьях 1980-х годов это была вполне приемлемая эмоция.

В существующих процедурах разметки текстовых данных (см. например [Francisco, Hervás, Gervás, 2007; Kolmogorova, Kalinin, Malikova, 2020]) фактор социальной желательности / нежелательности эмоции пока не принимался во внимание, что, в общем, приводит к недостаточно высокому качеству работы нейросетевых моделей, обученных на таких датасетах.

В данной публикации мы бы хотели представить результаты нашего пилотного эксперимента по разметке эмоциональных текстовых фрагментов, произведенной на основе концепции Байесовской сыворотки правды (*Bayesian Truth Serum*; далее – BTS). Разработанная для проведения опросов в тех сферах, где критерии

истинности неприменимы, концепция переносит фокус исследовательского внимания на категорию правдивости ответов информантов, позволяя выявить случаи несовпадения «личной правдивости» и «правдивости других». Благодаря текстам, при оценке которых такое несовпадение случается, мы можем обнаружить неявные правила эмоционального социального поведения, принятые в сообществе.

Байесовская сыворотка правды

Байесовская сыворотка правды – это метод, который позволяет работать с субъективной информацией, где абсолютная истина практически непознаваема, и при этом иметь возможность вознаграждать информантов за правдивость [Carvalho, Larson, 2011]. BTS используется для опросов с закрытыми вопросами. Участникам предлагается, помимо указания личного мнения (ответ на тот или иной вопрос), либо предположить, какой ответ выбрало большинство респондентов, либо оценить процент выбравших тот же ответ, что и участник опроса, или, например, дать процентное предсказание выборов каждого варианта ответа. Для повышения эффективности метода опрашиваемым предлагается награда за верное или близкое к истине предположение [Prelec, 2004].

При анализе собранных данных в первую очередь во внимание принимается разница личных ответов и предсказаний. Если данное значение невелико, выдвигается гипотеза о том, что респондент дал правдивый личный ответ, в противном случае предполагается, что выбор участника может быть предвзятым или намеренно неправдивым в силу влияния некоторого скрытого фактора социальной природы.

Таким образом, в рамках BTS-концепции мы получаем две важные категории ответов: 1) «неожиданно совпадающие ответы» (*surprisingly common answers*), когда большинство респондентов, находясь в позиции «предсказателя», спрогнозировали низкую частотность некоторого ответа, а в реальности доля информантов, ответивших так, была велика, то есть в точке пересечения ожидаемых информантами ответов и ответов, реально полученных в той же выборке, к которой принадлежали они сами и относительно которой строился прогноз, реальность превзошла ожидания; 2) «неожиданно несовпадающие ответы» (*surprisingly uncommon answers*), когда большинство респондентов, находясь в позиции

«предсказателей», спрогнозировали высокую частотность некоторого ответа в выборке, но в реальности этот ответ выбрали мало информантов, то есть в точке пересечения ожидаемых информантами ответов и ответов, реально полученных в выборке, ожидания превзошли реальность. И в первом, и во втором случае мы имеем дело с такими сферами социальных отношений, которые имеют скрытые правила и ограничения на публичное выражение и оценку эмоций. Получив в результате эмоциональной разметки согласно BTS-концепции некоторую выборку текстов, оценки по которым входят в категории «неожиданно совпадающих» или «неожиданно несовпадающих», мы можем продолжить анализировать данные тексты психологическими, социологическими и лингвистическими методами, чтобы больше узнать об этих неявных нормах эмоционального поведения в сообществе.

BTS широко используется во многих областях, например таких, как социология науки [The use of questionable …, 2021] и менеджмент [Applications of Bayesian …, 2021], однако для разметки эмоциональных данных он применяется, как свидетельствует проведенный нами анализ научной литературы, впервые.

Материал и методы

В данном исследовании эмоциональная разметка выборки текстов производится на базе модели PAD (*Pleasure-Arousal-Dominance model*) [Russel, Mehrabian, 1977]. Модель эмоционального состояния PAD разработана Альбертом Мехрабианом и Джеймсом А. Расселом для описания и измерения эмоциональных состояний.

Она относится к группе так называемых пространственных моделей эмоций[1]. Согласно модели PAD, любая эмоция или эмоциональное состояние могут быть обозначены как точка в трехмерном пространстве с осями *Pleasure* (удовольствие), *Arousal* (возбуждение) и *Dominance* (доминантность) (рис. 1). Первое измерение показывает, насколько эмоция приятна или неприятна

[1] Если категориальные модели эмоций П. Экмана, Р. Плутчика и С. Томкинса представляют эмоции как дискретные категории с четкими границами, то в рамках пространственных моделей PAD, VAD или Куб Лёвхема отдельное эмоциональное состояние рассматривается как точка в эмоциональном континууме, предопределяемом влиянием некоторых трех факторов психо- или нейрофизиологической природы (в зависимости от концепции).

человеку. Второе демонстрирует степень яркости переживания эмоции: например, любопытство или злость переживаются ярко, субъект эмоции находится в возбужденном состоянии, а скука, наоборот, предполагает некоторую заторможенность, замедленность всех психических реакций субъекта эмоции. Третий показатель выражает тот уровень контроля над ситуацией, над внешним стимулом эмоции, который ощущает субъект эмоции: проявления гнева и отвращения трудно поддаются эмоциональному контролю, а, например, любопытство достаточно легко сдержать [Russell, 1980; Mehrabian, 1996].

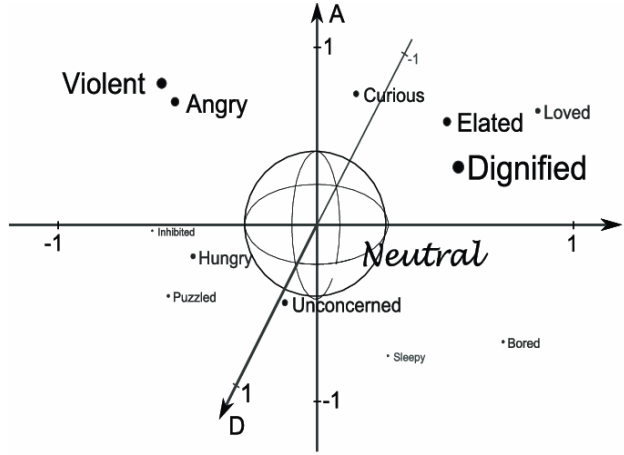

Рис. 1. Эмоциональная модель PAD [Mehrabian, 1996]:
P – *Pleasure*, A – *Arousal*, D – *Dominance*

Как правило, при использовании данной модели каждый показатель оценивается по шкале от -1 до 1. Но в нашей работе предполагался анализ значений отклонений, поэтому было принято решение использовать шкалу с более широким диапазоном: от 0 до 10.

Сбор датасета

В качестве основы датасета использовались тексты, собранные с помощью парсинга из группы «Подслушано» в социальной сети ВКонтакте. Выбор именно этого источника обусловлен тем, что он содержит регулярно публикуемые тексты высокой степени эмоциональности, написанные пользователями социальных интер-

нет-сетей в жанре так называемого «интернет-откровения». Каждый текст имеет тег, например #Подслушано_одиночество или #Подслушано_счастье и т.д. Проведя психолингвистический эксперимент, в котором информантам (N = 30, mean age = 19,4, носители русского языка, студенты гуманитарных специальностей) предлагались тексты (N = 50), размещенные под разными эмоциональными хештегами, и нужно было соотнести текст с одной из восьми эмоций (по С. Томкинсу) или отнести его к категории «нейтральное» (при этом информанты хештеги не видели).

Мы отобрали те хештеги, которые предваряли тексты, отнесенные простым большинством информантов к той или иной эмоции, как условные маркеры того или иного эмоционального класса текстов. Таким образом, мы создали пары «хештег – эмоция», например, #Подслушано_БЕСИТ – гнев, #Подслушано_стыдно – стыд. Далее все тексты, которые извлекались под данными хештегами, маркировались по умолчанию эмоциональной меткой, ассоциированной с данным хештегом. Таким образом, все тексты были распределены на группы, согласно восьмикомпонентной модели эмоций С. Томкинса [Tomkins, 2014]: страх, гнев, отвращение, грусть, радость, удивление, воодушевление, стыд (в англоязычном варианте С. Томкинса: *fear*, *anger*, *disgust*, *distress*, *enjoyment*, *startle*, *excitement*, *shame*). Русскоязычные аналоги англоязычных эмоций также были определены в психолингвистическом эксперименте с 30 информантами (mean age = 21,3; носители русского языка, высшее гуманитарное образование, 21 женщина, 9 мужчин). Информантам предъявляли фотографии лиц людей, использованные С. Томкинсом [Tomkins, 1981], и просили назвать от 1 до 3 слов русского языка, называющих эмоцию, испытываемую информантами. Наиболее часто повторяющиеся номинации выбирались в качестве эквивалентов и использовались в дальнейшем. Подчеркнем, что данный эксперимент носил предварительный характер и использовался исключительно для того, чтобы проверить адекватность адаптации модели эмоций С. Томкинса для русского языка.

Для экспериментальной разметки были отобраны 300 текстов. Помимо удаления хештегов, тексты перед составлением опроса никак не изменялись. Главным критерием отбора послужил размер текстов. Тексты слишком маленького размера (менее 80 токенов) могут быть неинформативны и являться нерелевантными для целей исследования, большие же тексты (более 200 токенов) не позволили бы респондентам удерживать в достаточной степени вни-

мание и концентрацию. Кроме того, мы постарались соблюсти, насколько это было возможно, гендерный баланс среди анонимных авторов текстов, опираясь на грамматические маркеры рода (например, я сказала, я сказал).

Таким образом, в итоговую выборку попали тексты, содержащие по 10–15 предложений (80–120 токенов). Для эмоций *радость*, *страх*, *удивление*, *отвращение* было отобрано по 37 текстов, для *воодушевления*, *гнева*, *стыда*, *грусти* – по 38 текстов. В каждой группе оказались тексты 18 авторов женского пола и такого же количества авторов мужского пола, а также по 1–2 текста, у которых пол автора не определяется в виду отсутствия эксплицитных маркеров рода. Иной информацией об авторах текстов мы не располагали, поскольку тексты извлекались нами из уже опубликованных во ВКонтакте – у нас не было задачи получить их экспериментальным путем.

Разметка проводилось на платформе Google Forms.

Было составлено 12 опросов, в каждом из которых представлено по 25 текстов. Чтобы опросники заполнялись равномерно и рандомизированно, мы сформировали дополнительный распределительный опрос, в котором участники выбирали число из выпадающего списка, после чего попадали на страницу со ссылкой на конкретную анкету (рис. 2).

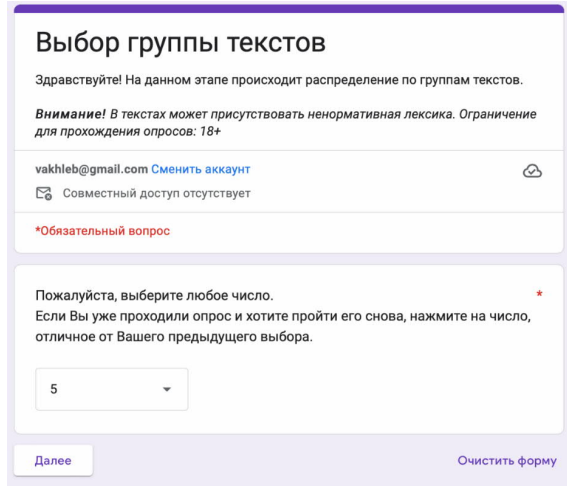

Рис. 2. Интерфейс оценщика на этапе выбора группы текстов

На приветственном экране каждого из 25 опросников информанты видели общее описание задачи: Здравствуйте! Предлагаем Вам принять участие в нашем исследовании. Вам будут последовательно предложены 25 коротких текстов (10–15 предложений) из социальной интернет-сети ВКонтакте. После прочтения каждого текста Вам необходимо будет оценить по шкале от 0 до 10 а) интенсивность «приятности» эмоций, которые автор текста испытал; б) степень яркости этих эмоций; в) насколько автор мог контролировать свои эмоции в описываемый в тексте момент.

Пример текста для оценки представлен на рис. 3.

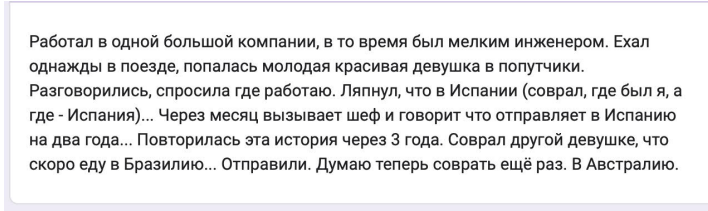

Рис. 3. Пример текста в опросе

Респондентам предлагалось по шесть закрытых вопросов к каждой истории, направленных на оценку эмоционального состояния автора текста: три вопроса о личном мнении участника об этом состоянии и три вопроса, предполагающих выдвижение предположения об ответах большинства на тот же вопрос:

1. Как Вы считаете, насколько приятны были автору истории описываемые переживания (отметьте нужный балл на шкале, где 0 значит «Автор испытал сильное неприятное чувство», а 10 – «Автору было крайне приятно»).

2. С Вашей точки зрения, насколько яркими были эмоции, испытанные автором текста в описываемой им ситуации (отметьте нужный балл на шкале, где 0 значит «Совсем неяркие, автор не был ни взбудоражен, ни возбужден», а 10 – «Очень яркие, с эффектом сильного возбуждения»).

3. По Вашему мнению, как сильно над автором доминировали испытанные эмоции, насколько он мог их контролировать (отметьте нужный балл на шкале, где 0 значит «Эмоции были под полным контролем», а 10 – «Эмоции невозможно было контролировать»)?

4. Как Вы думаете, какое значение на такой же шкале выбирает большинство респондентов, оценивая приятность эмоций, пережитых автором?

5. Как Вы думаете, какое значение на такой же шкале выбирает большинство респондентов, оценивая яркость эмоций, полученных автором истории?

6. Как Вы думаете, какое значение на такой же шкале выбирает большинство респондентов, оценивая доминантность эмоций, испытанных автором?

Рис. 4. Пример расположения шкалы в опросе

Под каждым вопросом была расположена измерительная одиннадцатибалльная шкала (от 0 до 10 (рис. 4)). Шкалы были составлены с использованием логики метода семантического дифференциала Ч. Осгуда [Osgood, 1969]. Минимальные и максимальные значения измерений маркированы и сопровождены дополнительными интерпретациями в форме утверждений.

Информантам сообщалось, что в случае, если их предсказание относительно самого частотного ответа совпадет хотя бы единожды при оценке 10 текстов, то они получат набор стикеров с логотипом лаборатории языковой конвергенции НИУ «Высшая школа экономики – Санкт-Петербург».

Каждый текст был размечен десятью информантами. Всего в опросе участвовало 120 информантов. Общая характеристика респондентов: жители Москвы или Санкт-Петербурга в возрасте от 18 до 55 лет (mean age = 29.3), 90 женщин и 30 мужчин, все – носители русского языка. Мы намеренно не контролировали такие переменные, как род занятий, возраст и образование в выборке информантов, поскольку предполагали, что гетерогенность выборки поможет выявить некоторые паттерны, присущие всему культурно-языковому

коллективу. Однако мы считаем, что значения данных переменных могут оказывать влияние на результаты оценки, поэтому в дальнейшем планируем провести эксперименты в разных выборках и сопоставить результаты.

Кроме того, что мы разработали и реализовали опрос, мы провели предобработку самих текстовых данных, предложенных информантам для разметки.

Для обработки данных использовались модули Re (Lib/re/) и String (Lib/string.py) для языка Python, а также библиотека Natasha (https://github.com/natasha/natasha). Модуль Re позволяет работать с текстовыми данными, используя регулярные выражения; встроенный модуль String предназначен для работы со строками. С помощью этих модулей тексты были разделены на отдельные предложения, а также очищены от знаков препинания и числовых переменных. Natasha – это одна из самых используемых библиотек для обработки текстовых данных на русском языке, содержащая инструменты и модели, обученные на корпусе новостных статей. Ресурсы библиотеки были использованы для токенизации и лемматизации данных.

Результаты и обсуждение

После выгрузки результатов эксперимента в формате таблиц с платформы Google Forms мы приступили к их анализу, который проводился на платформе Google Colaboratory. Google Colab – это облачная среда для разработки и выполнения программного кода на основе Jupyter Notebook. Для анализа собранного датасета использовались библиотеки Python (Pandas и NumPy), для визуализации – Matplotlib. Pandas (Panel Data) используется для обработки и анализа структурированных данных. Библиотека NumPy (Numeric Python) предназначена для выполнения математических вычислений, работы с многомерными массивами и матрицами. Matplotlib применяется для визуализации данных двумерной и трехмерной графикой.

Первоначально ответы респондентов были внесены в датафрейм. На рис. 5 продемонстрирован разброс оценок, полученных от десяти информантов для одного из текстов: красными точками отображены показатели личных ответов, зелеными – показаны значения предсказанных оценок.

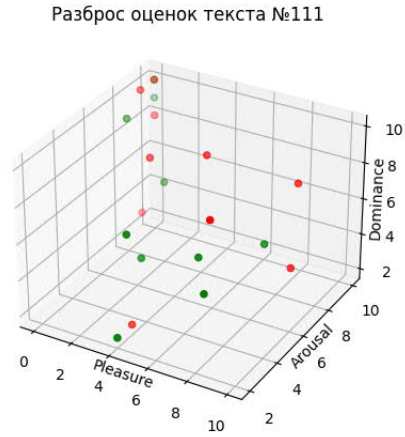

Рис. 5. Личные (красные точки) и предсказанные (зеленые) оценки для одного из текстов выборки

Можно предположить, что пары близких точек разного цвета – это личное мнение и ответ-предсказание одного и того же респондента. Единичные точки, не образующие заметных пар, представляют статистически значимые оценки – предсказанный ответ сильно отличается от собственного ответа.

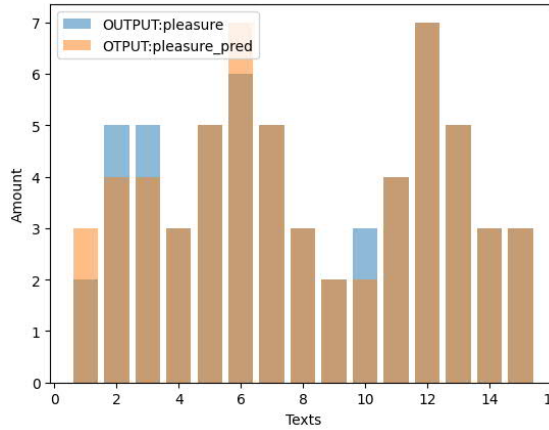

Рис. 6. Средние значения оценок по шкале *Pleasure* для личных оценок

После этого были посчитаны средние значения оценок показателей *pleasure*, *arousal*, *dominance* (личная оценка респондента) и pleasure_pred, arousal_pred, dominance_pred (предсказание). На рис. 6–8 приведены сравнения средних значений на каждой из трех шкал по личным оценкам и по предсказанным оценкам для первых 15 текстов датасета: голубым показаны отличия для личных оценок, оранжевым – для предсказанных.

Например, на рис. 6 мы видим, что для текста № 1 (по хештегу принадлежит эмоциональному классу «страх») среднее значение предсказанной оценки по шкале *Pleasure* выше, чем оно получилось на самом деле, то есть, возможно, существует какой-то скрытый социальный фактор, который помешал информантам дать правдивые личные ответы, но они спроецировали свой «градус правдивости» на «других», предсказав более высокие баллы. Например, тот факт, что рассказчицу сбила машина и она лежала в больнице, был, возможно, расценен информантами как некий социальный ограничитель на более или менее высокие баллы личной оценки по шкале «приятности» – информанты постарались сохранить лицо, но решили, что другие ответят более правдиво, то есть с более высоким баллом приятности:

Текст № 1: *Живу одна, вышла в магазин, и меня сбила машина. Очнулась уже в больнице. Нихрена не помнила из-за сильного удара головой. Ни телефона, ни документов с собой не было. Личность установить не получалось. И почему-то все решили, что мне 15. Врач открытым текстом сказал, что если не вспомню, и не найдутся родственники, придётся в детдом. Может, хотел напугать. Через четыре дня вспомнила, кто я, и что мне 21. Мама уже собиралась в розыск объявлять, благо подруги мои немного её успокаивали, сами искали*[1].

Для текстов № 2 (эмоция «стыд») и № 3 (эмоция «воодушевление») мы видим обратную ситуацию – значения реальных оценок на один пункт превосходят предсказанные оценки. Очевидно, какие-то скрытые факторы, связанные с социальным престижем, с самопрезентацией, например, стимулировали респондентов слегка завысить свои личные оценки, но при этом они посчитали, что остальные, будучи не столь чувствительны к данным факторам Х, ответят более правдиво:

[1] Здесь и далее в примерах сохранены авторские орфография и пунктуация.

Текст № 2: *Подошли милые парень с девушкой. Лица приветливые, но растерянные. Думаю – иностранцы. А они на языке жестов меня о чем-то спрашивают... Я растерялась. «Нет, – говорю, – не понимаю, извините». Так неловко стало. Ушла. А потом думаю, вот дура, надо было ручку с бумагой дать. Уже третий день стыдно*;

Текст № 3: *Знакомый решил откосить от армии, прикинулся дурачком: гнал перед комиссией лабуду. Через шесть лет пошел в автошколу. А тут психиатр ему засаду устроил – дурачок же. Пришлось объясняться. Собрали по его честь консилиум, усадили в центр комнаты и давай его «щекотать», мол, ты ж тут доказывал, что самый умный – повтори. Тесты всякие проходил. Даже в дурке два месяца заставили полежать под наблюдением. Геройства поубавилось – машина, мамой купленная, ждала. Заключение*: *не дурачок он, а долбо*б. Права получил.*

Можно предположить, что в тексте № 3 информанты увидели ситуацию «возвращения социальной справедливости» и, поскольку, видимо, это ситуация одобрительно оценивается членами сообщества в целом, информант не скрывал (а может быть и слегка преувеличил) тот факт, что детектировал в тексте некоторую приятность эмоционального состояния рассказчика.

На шкалах *Arousal* (рис. 7) и *Dominance* (рис. 8) мы видим интересные расхождения между личной оценкой и предсказанной для текста № 10 (эмоция «отвращение»):

Текст № 10: *Батя рассказал, что когда я был еще мелкий, в их доме жила женщина в трешке. У нее была работа, на жизнь не жаловалась, но вот с головой у нее было не в порядке. Рано утром, пока еще не вывезли мусор из баков, она шла на ближайшую помойку и рылась в ней. Тащила домой какие-то вещи, выброшенные продукты домой. Всегда уходила с полными пакетами с помойки. Думала, что все спят и никто ее не видит, но отец страдал бессоницей и наблюдал в окно, как она тащится от всякой тухлятины. Она живая до сих пор, дети от нее разбежались давно, а она уже погрязла в вонючем хламе.*

На обеих шкалах («возбуждение» и «доминантность») предсказанная оценка превышает личную (реальную). По-видимому, информанты по какой-то причине посчитали неприемлемым для себя лично считать эмоцию отвращения у нарратора более или менее яркой и слабо поддающейся контролю, но решили, что анонимные «другие» поставят те баллы, которые не осмелились поставить респонденты в личных оценках. Можно предположить, что в сообществе существуют некие скрытые правила, ограничивающие яркость проявления эмоции отвращения и предписывающие максимально их контролировать.

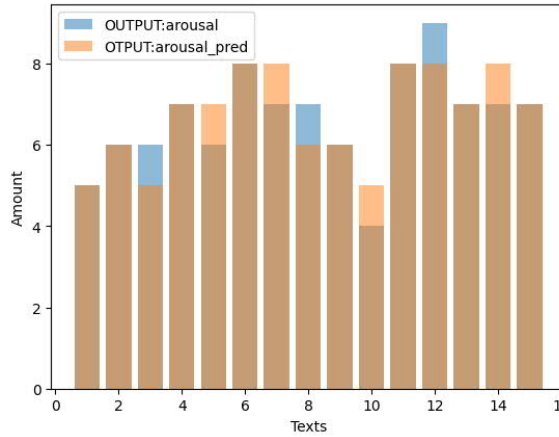

Рис. 7. Средние значения оценок по шкале *Arousal*

Далее с помощью библиотеки NumPy был произведен расчет стандартного отклонения личных и предсказанных оценок по каждой из трех шкал. Наша гипотеза состояла в том, что, предсказывая оценку, информанты будут иметь больший разброс в баллах, чем давая собственную оценку. Однако данная гипотеза подтвердилась только для шкалы *Pleasure* (рис. 9). Для остальных двух шкал средние значения стандартного отклонения практически не отличались.

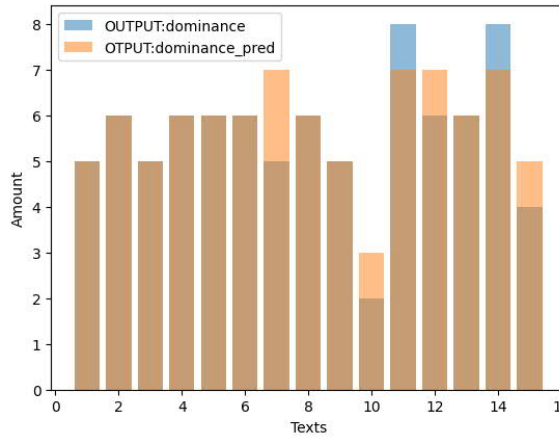

Рис. 8. Средние значения оценок по шкале *Dominance*

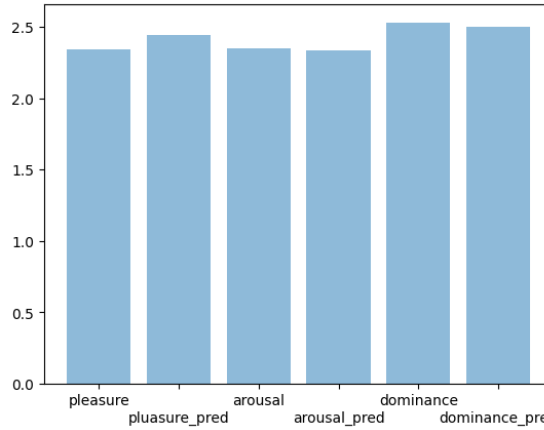

Рис. 9. Средние значения стандартных отклонений оценок

Далее мы отобрали тексты, для которых (по каждой шкале отдельно) личные и предсказанные оценки отличаются в наибольшей степени. В результате использования формулы (рис. 10) были сформированы три таблицы текстов с наибольшей разницей (≥2) между средними значениями личной и предсказанной оценок. Каждая из таблиц, полученных в результате вычислений, состоит из следующих столбцов данных: текст, порядковый номер текста (id), эмоция, к которой он принадлежит по хештегу, и значение разницы личной и предсказанной оценок.

```
def surprised_texts_of_state(state):

  surprised_common_emotions = []
  surprised_common_texts = []
  surprised_common_texts_ids = []
  surprised_common_dist = []

  l = 0

  for t in data[state]:
    if t >= 2:
      surprised_common_emotions.append(data['emotion'][l])
      surprised_common_texts.append(data['text'][l])
      surprised_common_texts_ids.append(data['text_id'][l])
      surprised_common_dist.append(data[state][l])
    l += 1

  df = pd.DataFrame({'text_id' : surprised_common_texts_ids,
                     'text' : surprised_common_texts,
                     'emotion' : surprised_common_emotions,
                     f'{state}' : surprised_common_dist,
                     })

  return df
```

Рис. 10. Формула составления таблицы с текстами, для которых значения личной и предсказанной оценок разнятся в наибольшей степени

Итак, по шкале *Pleasure* в таблицу попали 16 текстов, *Arousal* – 6, *Dominance* – 13 текстов. Поскольку у нас были данные о принадлежности текста тому или иному эмоциональному классу из восьми возможных (по С. Томкинсу) согласно парам «хештег – эмоция», мы посмотрели, тексты каких эмоциональных классов вошли в группу, спровоцировавшую «неожиданно совпадающие ответы» и «неожиданно несовпадающие ответы». На рис. 11–13 отображено процентное отношение эмоций для текстов, попавших в данную группу.

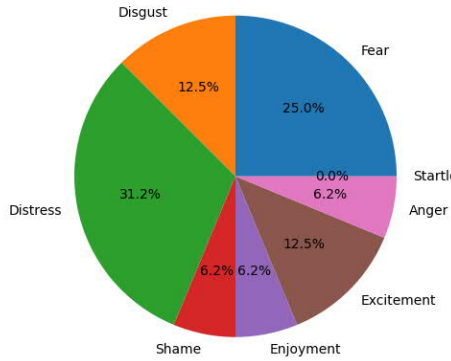

Рис. 11. Процентное отношение эмоций в группе «неожиданно совпадающих / несовпадающих ответов» по шкале Pleasure

По шкале *Pleasure* «неожиданные» ответы спровоцировали в основном тексты из классов «грусть» (*distress*), «страх» (*fear*).

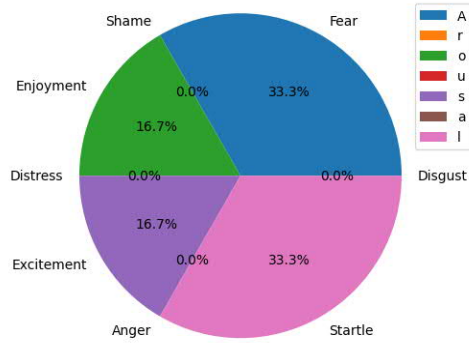

Рис. 12. Процентное отношение эмоций в группе «неожиданно совпадающих / несовпадающих ответов» по шкале *Arousal*

По шкале *Arousal* – тексты из эмоциональных классов «страх» (*fear*) и «удивление» (*startle*). По шкале *Dominance* – тексты из эмоциональных классов «отвращение» (*disgust*) и «воодушевление» (*excitement*).

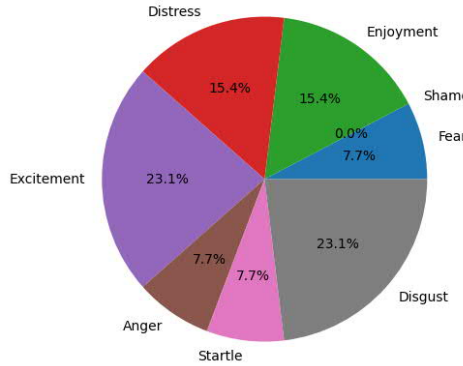

Рис. 13. Процентное отношение эмоций в группе «неожиданно совпадающих / несовпадающих ответов» по шкале *Dominance*

Для текстов, попавших в группы наибольшей разности личных и предсказанных оценок, после предварительной лемматизации и удаления стоп-слов мы составили облака слов (рис.14–16).

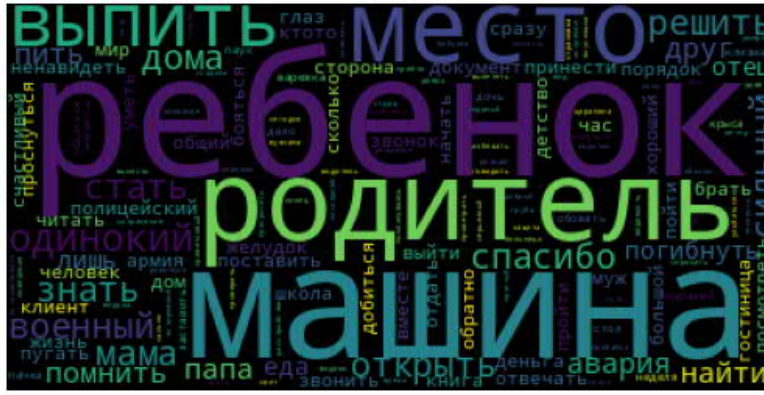

Рис. 14. Облако слов для текстов с наибольшей разницей в личных и предсказанных оценках (*Pleasure*)

Во всех трех «облаках» самыми частотными являются существительные, называющие субъектов семейных (родитель, ребенок, мама, муж) и личных (парень, друг) отношений.

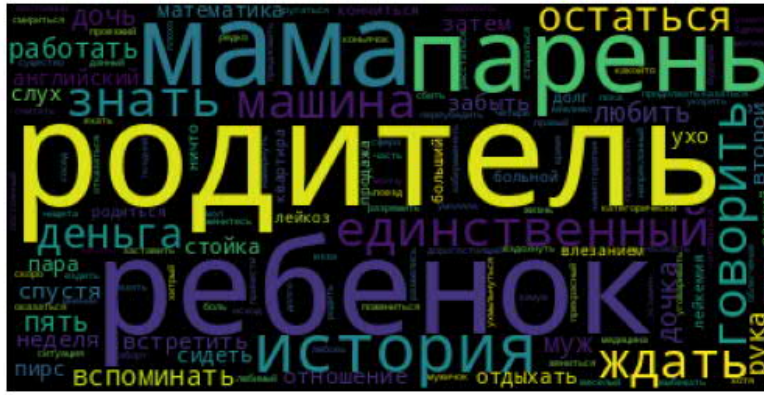

Рис. 15. Облако слов для текстов с наибольшей разницей в личных и предсказанных оценках (*Arousal*)

Рис. 16. Облако слов для текстов с наибольшей разницей в личных и предсказанных оценках (*Dominance*)

Однако есть и отличия. Для шкалы *Dominance* наиболее частотными являются номинации друг и муж (родитель – лемма для частотной во всех группах словоформы родители), а также глаголы, имеющие преимущественно семантику перемещения в простран-

стве: остаться, сидеть, выйти, уехать, вернуться. Для шкалы *Arousal* частотными являются лексические единицы, связанные с ситуацией рассказывания истории о материнстве: мама, (единственный) ребенок, говорить, ждать, вспоминать. Для *Pleasure*, по-видимому, характерной ситуацией, провоцирующей неожиданно совпадающие и неожиданно несовпадающие ответы, является нечто, связанное с безопасностью на дороге и семьей: машина, выпить, ребенок, родитель.

Полагаем, что в нашей выборке определились три наиболее регламентируемые неявными социальными правилами траектории эмоционального поведения в рамках того сообщества, к которому принадлежат наши информанты – жители мегаполисов (Москва и Санкт-Петербург). Это ситуации, затрагивающие отношения в диадах «муж (парень) – жена (девушка)» и «мать – ребенок», а также случаи девиантного поведения (выпить и сесть за руль), связанные с безопасностью на дороге, в том числе – с безопасностью членов семьи.

Кроме того, отметим, что интерес представляют и тексты, для которых личные оценки совпали с предсказанными. Мы рассматриваем их как валидных кандидатов в обучающую выборку для дообучения русскоязычных моделей, способных решать задачу автоматического эмоционального анализа.

Заключение

Проведенный пилотный эксперимент по применению BTS-концепции для разметки эмоциональных текстов в целом демонстрирует интересный потенциал данной методологии: 1) с ее помощью можно получать достаточно достоверные данные для дальнейшего обучения моделей (тексты, для которых значения личной и предсказанной оценки совпали или близки); 2) на ее основе мы можем получать сеты текстов, отражающих ситуации, на которые распространяются некие неявные правила и ограничения, регламентирующие в рамках данного социума эмоциональное поведение его членов. С последними текстами мы можем продолжать работу в рамках междисциплинарного исследования, поставив цель выявления таких скрытых паттернов эмоционального поведения.

Список литературы

Applications of Bayesian approaches in construction management research: a systematic review / Hon C.K., Sun C., Xia B., Jimmieson N.L., Way K.A., Wu P.P.-Y. // Eng. Const. Arch. Manag. – 2022. – Vol. 29., No.5. – P. 2153–2182. DOI: https://doi.org/10.1108/ECAM-10-2020-0817

Carvalho A., *Larson K.A.* Truth Serum for Sharing Rewards // 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6. – 2011. − Volume 1-3. – P. 635-642.

Charaudeau P. Identité sociale et identité discursive. Un jeu de miroir fondateur de l'activité langagière // Identités sociales et discursives du sujet parlant / Charaudeau P. (dir.). – Paris : L'Harmattan, 2009. – P. 15−27.

Charaudeau P. La pathémisation à la télévision comme stratégie d'authenticité // Les émotions dans les interactions. – Lyon : Presses universitaires de Lyon, 2000. – P. 125−156.

Francisco V., *Hervás R.*, *Gervás P.* Two different approaches to automated mark up of emotions in text // Research and Development in Intelligent Systems XXIII, SGAI 2006 / Bramer, M., Coenen, F., Tuson, A. (eds.). – London : Springer, 2007. – P. 101−114. DOI: https://doi.org/10.1007/978-1-84628-663-6_8

Kolmogorova A., *Kalinin A.*, *Malikova A.* Non-discrete sentiment dataset annotation: case-study for Lövheim Cube emotional model // Communications in Computer and Information Science. – 2020. – Vol. 1242. – P. 154−164.

Mehrabian A. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament // Curr. Psychol. – 1996. – Vol. 14. – P. 261–292. DOI: 10.1007/BF02686918

Osgood C.E. On the whys and wherefores of E.P // Journal of personality and social psychology. – 1969. – Vol. 123. – P. 194–199.

Prelec D. A Bayesian truth serum for subjective data // Science. – 2004. – Vol. 306 (5695). – P. 462–466.

Russell J.A. A circumplex model of affect // J. Personal. Soc. Psychol. – 1980. – Vol. 39. – P. 1161–1178. DOI: https://doi.org/10.1037/h0077714

Russell J.A., *Mehrabian A.* Evidence for a three-factor theory of emotions // Journal of Research in Personality. – 1977. – Vol. 11(3). – P. 273–294.

The use of questionable research practices to survive in academia examined with expert elicitation, prior-data conflicts, Bayes factors for replication effects, and the Bayes Truth Serum / van de Schoot R., Winter S.D., Griffioen E., Grimmelikhuijsen S., Arts I., Veen D., Grandfield E.M., Tummers L.G. // Front. Psychol. – 2021. – Vol. 12. – P. 621547. DOI: https://doi.org/10.3389/fpsyg.2021.621547

Tomkins S. Affect Theory // Approaches to emotion. – New York : Psychology Press, 2014. – P. 163−195.

Tomkins S. The quest for primary motives: biography and autobiography of an idea // Journal of Personality and Social Psychology. – 1981. – Vol. 41(2). – P. 308–329. DOI: https://doi.org/10.1037/0022-3514.41.2.306

References

Carvalho, A., Larson, K. (2011). A Truth Serum for Sharing Rewards. *10th International Conference on Autonomous Agents and Multiagent Systems* (*AAMAS 2011*), Taipei, Taiwan, May 2–6, 1–3, 635–642.

Charaudeau, P. (2000). La pathémisation à la télévision comme stratégie d'authenticité. In *Les émotions dans les interactions* (pp. 125–156). Lyon: Presses universitaires de Lyon.

Charaudeau, P. (2009). Identité sociale et identité discursive. Un jeu de miroir fondateur de l'activité langagière, in Charaudeau P. (dir.). In *Identités sociales et discursives du sujet parlant* (pp. 15–27). Paris: L'Harmattan.

Francisco, V., Hervás, R., Gervás, P. (2007). Two different approaches to automated mark up of emotions in text. In Bramer, M., Coenen, F., Tuson, A. (eds.) *Research and Development in Intelligent Systems XXIII*, *SGAI 2006*. (pp. 101–114). London: Springer. DOI: https://doi.org/10.1007/978-1-84628-663-6_8

Hon, C.K., Sun, C., Xia, B., Jimmieson, N.L., Way, K.A., Wu, P.P.-Y. (2022). Applications of Bayesian approaches in construction management research: a systematic review. *Eng. Const. Arch. Manag*, *29*, no.5, 2153–2182. DOI: https://doi.org/10.1108/ECAM-10-2020-0817

Kolmogorova, A., Kalinin, A., Malikova, A. (2020). Non-discrete sentiment dataset annotation: case-study for Lövheim Cube emotional model. *Communications in Computer and Information Science*, *1242*, 154–164.

Mehrabian, A. (1996). Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr. Psychol.*, *14*, 261–292. DOI: 10.1007/BF02686918

Osgood, C.E. (1969). On the whys and wherefores of E.P. *Journal of personality and social psychology, 123*, 194–199.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(*5695*), 462–466.

Russell, J.A. (1980). A circumplex model of affect. *J. Personal. Soc. Psychol.*, *39*, 1161–1178. DOI: https://doi.org/10.1037/h0077714

Russell, J.A., Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(*3*), 273–294.

van de Schoot, R., Winter, S.D., Griffioen, E., Grimmelikhuijsen, S., Arts, I., Veen, D., Grandfield, E.M., Tummers, L.G. (2021). The use of questionable research practices to survive in academia examined with expert elicitation, prior-data conflicts, Bayes factors for replication effects, and the BayesTruth Serum. *Front. Psychol.*, *12*, 621547. DOI: https://doi.org/10.3389/fpsyg.2021.621547

Tomkins, S. (2014). Affect Theory. In *Approaches to emotion* (pp.163–195). New York: Psychology Press.

Tomkins, S. (1981). The quest for primary motives: biography and autobiography of an idea. *Journal of Personality and Social Psychology*, *41*(*2*), 308–329. DOI: https://doi.org/10.1037/0022-3514.41.2.306

Об авторах

Колмогорова Анастасия Владимировна – доктор филологических наук, профессор, заведующий лабораторией языковой конвергенции, НИУ «Высшая школа экономики – Санкт-Петербург», Россия, Санкт-Петербург, akolmogorova@hse.ru

Хлебникова Василиса Андреевна – младший специалист тестирования, Общество с ограниченной ответственностью «Яндекс Крауд», Россия, Москва, va.khleb@yandex.ru

About the authors

Kolmogorova Anastasia Vladimirovna – Doctor of Philology, Professor, Head of the Laboratory of Language Convergence, Higher School of Economics – Saint-Petersburg, Russia, Saint-Petersburg, akolmogorova@hse.ru

Khlebnikova Vasilisa Andreevna – Junior Testing Specialist, Yandex Crowd Limited Liability Company, Russia, Moscow, va.khleb@yandex.ru