НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2023 Том 23 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2023 Vol. 23 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-2-279-288
УДК 004.8 004.056.53

Предсказание результатов 16-факторного теста Р. Кеттелла на основе анализа текстовых постов пользователей социальной сети

Валерий Дмитриевич Олисеенко[1], Максим Викторович Абрамов[2]✉

[1,2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] vdo@dscs.pro, https://orcid.org/0000-0002-3479-0085
[2] mva@dscs.pro✉, https://orcid.org/0000-0002-5476-3025

Аннотация

Предмет исследования. Исследована возможность автоматизации предсказания по небольшому набору данных оценки выраженности психологических особенностей по 16-факторному личностному тесту Р. Кеттелла пользователей социальной сети на основе анализа публикуемых ими на своей странице текстовых постов. **Метод.** Предложенный новый метод автоматизации оценки выраженности психологических особенностей по 16-факторному личностному тесту Р. Кеттелла включает в себя языковые модели и нейронные сети. Реализация метода предусматривает несколько шагов. На первом шаге происходит извлечение из аккаунтов пользователей социальной сети текстовых постов, их предобработка с помощью языковой модели RuBERT и ранее обученной достроенной над ней полносвязной нейронной сети. Итогом этого шага является нормализованное эмпирическое распределение постов по ранее введенным классам по каждому пользователю. Впоследствии на основе распределения постов пользователей производится оценка выраженности психологических особенностей пользователя с использованием метода опорных векторов, случайного леса и наивного байесовского классификатора. **Основные результаты.** Финальный набор данных для построения моделей и дальнейшего тестирования их работы составлен из 183 респондентов, прошедших тест Р. Кеттелла, со ссылками на их открытые аккаунты в социальной сети. Построены классификаторы, предсказывающие результаты для шести факторов (A, B, F, I, N, Q1) 16-факторного личностного теста Р. Кеттелла. **Практическая значимость.** Полученные результаты могут найти применение при создании прототипа автоматизированной системы предсказания оценки выраженности психологических особенностей пользователей социальной сети. Результаты работы полезны в прикладных и исследовательских системах, связанных с маркетингом, психологией и социологией, а также в области защиты пользователей от социоинженерных атак.

Ключевые слова

социальные сети, классификация текстов, искусственный интеллект, 16-факторный тест Р. Кеттелла, машинное обучение, нейронные сети

Благодарности

Работа выполнена в рамках проекта по государственному заданию СПБ ФИЦ РАН № FFZF-2022-0003; при финансовой поддержке РФФИ, проект № 20-07-00839; при финансовой поддержке гранта Президента Российской Федерации МК5237.2022.1.6.

Ссылка для цитирования: Олисеенко В.Д., Абрамов М.В. Предсказание результатов 16-факторного теста Р. Кеттелла на основе анализа текстовых постов пользователей социальной сети // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 2. С. 279–288. doi: 10.17586/2226-1494-2023-23-2-279-288

Predicting the results of the 16-factor R. Cattell test based on the analysis of text posts of social network users

Valerii D. Oliseenko[1], Maxim V. Abramov[2]✉

[1,2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] vdo@dscs.pro, https://orcid.org/0000-0002-3479-0085
[2] mva@dscs.pro✉, https://orcid.org/0000-0002-5476-3025

Abstract

We investigated the possibility of automating the prediction of the 16-factor personality traits by R. Cattell from text posts of social media users. The proposed new method of automating the evaluation of R. Kettell's 16-factor personality test traits includes language models and neural networks. Implementation of the method involves several steps. At the first step text posts are extracted from user accounts of social media, pre-processed with language model RuBERT and previously trained over a full-connected neural network. The result of this step is a normalized empirical distribution of the posts by the previously introduced classes for each user. Subsequently, based on the distribution of user posts the evaluation of the expression of psychological features of the user is made with the help of support vector machine, random forest and Naive Bayesian classifier. The final data set for model building and further testing their performance was made up of 183 respondents who took the R. Cattell test, with links to their public social media accounts. Classifiers predicting results for six factors (A, B, F, I, N, Q1) of R. Cattells 16-factor personality test were constructed. The results can be used to create a prototype of automated system for predicting the severity of psychological features of social media users. Results of work are useful in the applied and research systems connected with marketing, psychology and sociology, and also in the field of protection of users from social engineering attacks.

Keywords

online social networks, text classification, artificial intelligence, sixteen personality factor questionnaire, machine learning, neural networks

Acknowledgements

The research was carried out in the framework of the project on state assignment SPC RAS No. FFZF-2022-0003, with the financial support of the RFBR (No. 20-07-00839), with the financial support of the grant of the President of the Russian Federation MK 5237.2022.1.6.

For citation: Oliseenko V.D., Abramov M.V. Predicting the results of the 16-factor R. Cattell test based on the analysis of text posts of social network users. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 2, pp. 279–288. doi: 10.17586/2226-1494-2023-23-2-279-288

Введение

Вопрос оценки выраженности психологических особенностей пользователей социальных сетей встает во многих областях исследований: маркетинговых [1, 2], управления персоналом [3, 4], личностно-ориентированного обучения [5, 6], прогнозирования социального поведения [7–9], анализа защищенности персонала компаний от социоинженерных атак [10] и др. Данная оценка может проводиться при помощи: тестов-опросников, психодиагностических интервью, рисуночных тестов, ролевых игр, проективных методик и др. Однако процесс такой оценки часто затруднен сложностью ее проведения, как с точки зрения затрат временных и иных ресурсов, так и с позиции необходимости наличия компетентных специалистов или денежных ресурсов. Как отмечают некоторые исследователи [11], оптимальными для такой оценки являются тесты-опросники, так как они наиболее точны, валидны и надежны при умеренности требований к необходимому ресурсу. Вместе с тем психологические тесты можно реализовать в электронном формате, что сильно упрощает сбор и обработку данных в сравнении с другими подходами к оценке выраженности психологических особенностей пользователей, делает более оперативной выдачу результата.

Среди устоявшихся тестов для оценки выраженности психологических особенностей пользователей можно выделить следующие: «Большая пятерка» [12], методики ценностных ориентаций Ш. Шварца [13], 16-факторный личностный тест Р. Кеттелла [14], «Индекс жизненного стиля» Келлермана–Плутчика [15]. Данные тесты состоят из множества вопросов с несколькими альтернативными ответами. На основе результатов, полученных по итогам прохождения тестов, формируются порядковые шкалы для конкретной исследуемой черты личности. Результаты можно прогнозировать посредством применения методов искусственного интеллекта, нечеткой логики, математической статистики и т. п. к данным, извлекаемым из контента, публикуемого пользователя в социальных сетях [16, 17].

В работах [18, 19] рассмотрена автоматизация прогнозирования оценки выраженности личностных особенностей пользователя, на основе публикуемого им в социальных сетях контента. Предложен подход классификации текстовых постов пользователей по трем классам (подклассам): информационные (формальные, событийные, личные, интеллектуально-рассудительные, ссылочные, кулинарные); эмоциональные (позитивные, негативные и поздравительные); побудительно-деятельностные (благотворительные, продающие, побудительные к действию).

Классификация текстовых постов ранее производилась силами экспертов вручную, а в последствии была автоматизирована при помощи языковой модели

RuBERT[1] и достроенной над ней полносвязной нейронной сети [19]. Была также выявлена статистическая связь (корреляция) эмпирического распределения классифицированных постов (классов и подклассов) с оценкой выраженности психологических особенностей, полученных при помощи 16-факторного личностного теста Р. Кеттелла [14]. Необходимо на основе полученных результатов построить модели и алгоритмы, позволяющие прогнозировать оценки выраженности личностных особенностей пользователей на основе публикуемых ими в социальных сетях текстовых постов.

Цель работы — исследование возможности автоматизации по небольшому набору данных предсказания оценки выраженности психологических особенностей по 16-факторному личностному тесту Р. Кеттелла пользователей социальной сети на основе анализа публикуемых ими на своей странице текстовых постов. Теоретическая значимость заключается в разработке и проверке нового подхода, включающего в себя языковые модели и нейронные сети, который позволит автоматизировать процесс оценки выраженности личностных особенностей пользователей социальной сети. Практическая значимость состоит в создании наработок для прототипа автоматизированной системы предсказания оценки выраженности психологических особенностей пользователей социальной сети.

Релевантные работы

Среди существующих подходов к оценке выраженности психологических особенностей пользователей по их постам в социальных сетях можно выделить ручной и автоматизированный [20]. При ручном подходе эксперты (психологи) просматривают содержание личных страниц в социальной сети (или текст, извлеченный из них) и заполняют анкету с вопросом о личности данного пользователя. В автоматизированном подходе используются программы (в том числе с элементами искусственного интеллекта), которые анализируют текст, извлеченный из социальных сетей, и соотносят полученную информацию с результатами тестов, пройденных пользователями ранее. В работе [20] проведен мета-анализ статей, в которых рассмотрены подходы к анализу текста со страниц пользователей в социальных сетях для предсказывания результатов теста «Большая пятерка». В среднем полностью автоматизированные подходы показывают лучшие результаты по сравнению с экспертными, но их оценка может иметь сильное смещение в сторону заведомо ложных результатов тестов (например, из-за эффекта переобучения), не иметь интерпретации, понятной для эксперта. Предложенный в [16] подход на основе классификации постов изначально содержит экспертные оценки для классификации самих постов, что может нивелировать указанные выше недостатки.

Отдельная группа подходов к автоматизации оценки выраженности личностных особенностей пользователей опирается на использование открытого и закрытого словарей [21]. Закрытый словарь представляет собой список слов, которые можно отнести к лексике той или иной категории (например, к негативной категории относились слова «ненавижу», «плохой», «отвращение» и др.). Словарь называется закрытым, так как категории и слова, относящиеся к ним, не являются общедоступными, а предоставляются в составе каких-либо пакетов программ. Количество/распределение таких слов в тексте можно использовать для статистического анализа и предсказания класса текста (например, по жанрам, авторам, настроению текста и т. д.). Одна из таких систем — Linguistic Inquiry and Word Count [22, 23]. Однако проблема закрытых словарей заключалась в их реализации в платном программном обеспечении и частом отсутствии прямого доступа. Открытый словарь, наоборот, использует открытые данные, например, новостные сайты, Wikipedia, сообщения в социальных сетях и другие источники для поиска некоторых скрытых взаимосвязей (контекста) использования каждого слова в предложении [24]. Такой поиск происходит в том числе с применением методов искусственного интеллекта через поиск нестатистических взаимосвязей, что позволяет их использовать во множестве задач из области обработки естественных языков: классификация текста, создание чат-ботов, оцифровка речи и т. д. Среди подходов с открытыми словарями наиболее продвинутыми являются подходы на основе предобученных языковых моделей BERT [25], ElMo [26], OpenAI GPT-3 [27].

В работе [16] приведена схема классификации постов, в которой использован экспертный подход для разметки постов, а в [19] данный подход автоматизирован при помощи языковой модели RuBERT[1]. В научных работах часто используются исходные эмбеддинги [28, 29] языковых моделей или их комбинации с надстройкой некоторых моделей (например, нейронных сетей) над ними для «прямого» без построения промежуточного распределения по классам постов предсказания результатов психологического тестирования. Однако такие подходы могут быть лишены какой-либо интерпретируемости из-за их устройства. Вместе с тем в дальнейших исследованиях планируется рассмотреть такие «прямые» подходы на расширенном наборе данных и сравнить с результатами, получаемыми при помощи моделей, разработанных в настоящей работе.

Постановка задачи

16-факторный личностной тест Р. Кеттелла состоит из 187 вопросов с тремя вариантами ответа в каждом. Например, один из вопросов следующий: «Строя планы на будущее, я часто рассчитываю на простое везение», варианты ответов: «да/затрудняюсь ответить/нет». Результаты прохождения (ответы на вопросы) теста переводятся при помощи ключа в оценку на основе шкалы стенов — оценку со значением из целочисленного интервала [1; 10] по каждому из 16 факторов. Всего

[1] RuBert — это оригинальная модель BERT [18], обученная на основе русскоязычной Wikipedia [Электронный ресурс]. URL: https://habr.com/ru/company/sberbank/blog/567776/ (дата обращения: 22.02.2023).

в тесте оцениваются 16 следующих факторов: A — Общительность; B — Интеллект; C — Эмоциональная стабильность; E — Доминантность; F — Экспрессивность; G — Нормативность поведения; H — Смелость; I — Чувствительность; L — Подозрительность; M — Мечтательность; N — Дипломатичность; O — Тревожность; Q1 — Консерватизм; Q2 — Конформизм; Q3 — Самоконтроль; Q4 — Напряженность.

Таким образом, после прохождения опроса респондент получает оценку от 1 до 10 по каждому из 16 перечисленных факторов. Например, A (общительность) — 8, B (интеллект) — 10 и т. д. Далее выполняется сокращение множества значений оценки выраженности этих факторов следующим образом: если первоначальная оценка лежит на отрезке [1; 4], то она заменяется на значение «–1»; на отрезке [5; 6] — на «0»; на отрезке [7; 10] — на «1».

Такое сокращение множества значений оценки по каждому из 16 факторов: допустимо с точки зрения предметной области, в связи с тем, что психологов интересует наличие выраженности фактора в одну или другую сторону шкалы (т. е. бинарная оценка: сильно/слабо выражено); снижает размерность предсказываемой величины (метки класса), что позволяет строить предсказательные модели на малых данных. При помощи этого мы сводим задачу предсказания десяти категорий к трем. В дальнейшем такую оценку тоже можно производить, но на текущем этапе ограничимся тремя классами.

В работе [16] изучена взаимосвязь между количеством постов определенного класса (информационные; эмоциональные; побудительно-деятельностные) и выраженностью некоторых факторов. Например, фактор I (Чувствительность) положительно связан с количеством негативных постов ($r = 0{,}28$; $p < 0{,}05$[1]), факторы M и Q1 — информационных постов ($r = 0{,}25$; $p < 0{,}05$), факторы E и F — событийных постов ($r = 0{,}22$; $p < 0{,}05$ и $r = 0{,}22$; $p < 0{,}05$) и т. д.

В результате сформулируем задачу настоящей работы следующим образом. Необходимо по идентификатору пользователя в социальной сети «ВКонтакте» построить оценку выраженности его личностных особенностей по 16-факторному тесту Р. Кеттелла. Другими словами, предсказать, какие результаты может получить пользователь, если бы сам проходил тест.

В итоге задачу автоматизации предсказывания результатов сведем к 16 задачам мультиклассовой классификации, где меткой класса будет выступать категорированное значение результата одного из факторов (значение «–1» для оценки в интервале от 1 до 4; «0» — от 5 до 6; «1» — от 7 до 10). На вход опросника поступает общее число постов пользователя и значение числа постов по каждому классу, согласно классификации текстовых постов [17, 19].

Например, пусть пользователь опубликовал 5 эмоциональных текстовых постов, 3 — информационных, 2 — побудительно-деятельностных, всего постов — 10. Тогда на вход алгоритм получит следующие значения: эмоциональные — 0,5; информационные — 0,3; побудительно-деятельностные — 0,2. На выходе модели пользователь получит оценки степени выраженности факторов по тесту Р. Кеттелла.

Формирование и описание набора данных, метрики

Рассмотрим набор данных, полученный в результате исследования групп студентов высших учебных заведений Санкт-Петербурга на протяжении 2018–2020 гг. Исходный набор данных содержит 312 записей, каждая из которых представляет результат прохождения 16-факторного личностного опросника Р. Кеттелла и ссылку на страницу (ID) респондента в социальной сети «ВКонтакте»[2]. В дальнейшем данные из аккаунтов социальных сетей участников опроса выгружаются автоматизировано, с помощью метода VK API[3], и посты со всех страниц, которые удалось выгрузить. В социальной сети «ВКонтакте» пользователь может технически ограничить доступ к текстовым постам, публикуемым на своей странице, через настройки приватности, в этом случае автоматизировано выгрузить посты невозможно. Из исходного набора данных были исключены пользователи, которые ограничили настройками приватности доступ к своим постам, а также пользователи, не опубликовавшие ни одного поста. Данные исключения связаны с тем, что в текущем исследовании сделан фокус на построении оценок степени выраженности факторов на основе распределения текстовых постов в соответствии с классификацией. Ограничение пользователем доступа к своей странице или отсутствие публикаций на ней может также давать информацию о выраженности его личностных особенностей, что впоследствии планируется изучить и учитывать при анализе. В результате перечисленных исключений в наборе данных осталось 183 пользователя.

Процесс построения моделей для классификации постов подробно описан в работе [19]. Для новых неразмеченных данных логика моделей представлена следующим образом:

1) на вход системы подается ID пользователя в социальной сети «ВКонтакте»;
2) по указанному ID пользователя, если его страница открыта, выгружаются все текстовые посты;
3) производится предобработка текста (удаление стоп-слов, неинформативных символов, лемматизация);
4) предобработанный текст направляется в модель RuBERT[4] для получения эмбеддингов (векторного представления поста);

[1] r — коэффициент корреляции Пирсона и p — статистическая величина, используемая для проверки гипотезы на основе наблюдаемых данных.

[2] Все участвовавшие в исследовании респонденты дали соответствующее согласие на участие в исследовании и обработку обезличенных данных, полученных с их страниц в социальной сети.

[3] [Электронный ресурс]. URL: https://dev.vk.com/method (дата обращения: 22.02.2023).

[4] RuBert — это оригинальная модель BERT [18], обученная на основе русскоязычной Wikipedia [Электронный ресурс]. URL: https://habr.com/ru/company/sberbank/blog/567776/ (дата обращения: 22.02.2023).

5) полученные эмбеддинги пропускаются через три (по количеству классов) предобученные полносвязные четырехслойные нейронные сети для классификации их по подклассам в каждом классе.

Данные нейронные сети имеют четыре слоя со следующими характеристиками: первые три слоя функцию активации Relu (100, 50, 10 нейронов на каждом слое), четвертый — Softmax. Между первым и вторым слоями используется Dropout с коэффициентом 0,2 (функция выключения нейронов) для снижения эффекта переобучения. Функция потерь — категориальная кроссэнтропия. В процессе предобработки и дальнейшего эксперимента использован язык программирования Python 3.7.10 с библиотеками: NLTK[1] (для удаления стоп-слов), Re[2] (для удаления неинформативных символов, знаков, слов и т. д.), Pymorphy2[3] (для лемматизации слов). В качестве основы модели RuBERT применен фреймворк DeepPavlov[4], а для построения полносвязной нейронной сети — TensorFlow[5]. Валидация моделей проведена при помощи метода скользящего контроля по четырем тестовым блокам. Усредненные метрики на тестовых блоках показали следующие результаты для метрик F1-micro и F1-macro по классам: 0,748/0,698 (информационный); 0,894/0,840 (эмоциональный); 0,949/0,693 (побудительно-деятельностный). Отметим, что данные модели обучены на наборе данных, который размечен при помощи краудсорсингового проекта Yandex.Toloka. В табл. 1 показан фрагмент результата работы модели, построенной в [19].

Если пост не относится к классу, то в строке будет размещено значение с «No»: «No emotion»; «No information»; «No motivation». Таким образом, в первой строке с индексом 0 содержится информация о посте, опубликованном пользователем с ID = 10033***, который отнесен к подклассу позитивных (Positive) эмоционального класса, подклассу личных (Personal) информационного класса и подклассу продающих (Selling) побудительно-деятельностного класса. Отметим, что некоторые посты (например, пост с индексом 1 в табл. 1) модель отнесла сразу к не эмоциональному, не информационному и не побудительно-деятельностному. Далее такие посты удаляются из набора данных, так как результат свидетельствует об ошибке при классификации вследствие недостаточной обученности для корректной классификации. Изначально в наборе данных от 183 пользователей было 27 036 текстовых постов. После очистки от нераспределенных на классы постов осталось 25 283 поста. В табл. 2 представлен фрагмент финального набора данных по пользователям. Он состоит из 183 строк, каждая из которых соответствует конкретному пользователю, всего в наборе учтено 25 283 поста.

В табл. 2 столбцы A–Q4 представляют собой результаты оценки 16 факторов, полученных из пройденного теста 16-факторного личностного опросника Р. Кеттелла, преобразованные в соответствии с методикой, приведенной в разделе «Постановка задачи» (значения результатов фактора переведены по следующей схеме: от 1 до 4 — «–1»; от 5 до 6 — «0»; от 7 до 10 — «1»). В столбце ID представлены ID пользователей; fPost — общее количество постов пользователей; dPost — количество постов после удаления «неинформативных» постов, не содержащих текст или содержащих только неинформативные символы; столбцы Emotion, Information, Motivation — нормализованное на dPost количество постов каждого класса. Нормализация требуется, поскольку у разных пользователей разное количество постов (наибольшее число в наборе данных — 2515, наименьшее — 3), что потенциально может повлиять на выявление некорректных зависимостей обучаемой модели. Так как у разных пользователей количество постов может отличаться на несколько порядков, поэтому важно нормализовать соотношение у пользователя объема каждого класса к общему объему множества постов пользователя. Если нормализацию не выполнить, то это вызовет сложности на этапе обучения.

Рассмотрим пример работы модели для двух пользователей:

— пользователь 1 опубликовал: 5 эмоциональных текстовых постов, 3 — информационных, 2 — побудительно-деятельностных, всего постов — 10;
— пользователь 2 опубликовал: 10 эмоциональных текстовых постов, 6 — информационных, 4 — побудительно-деятельностных, всего постов — 20.

Несмотря на разное число публикованных постов, соотношение между классами сохраняется. Согласно [16], данное соотношение свидетельствует о выраженности тех или иных особенностей. Вместе с тем общее число постов пользователя также может нести в себе информацию о выраженности отдельных его особенностей, что рассматривается отдельно. Первичная описательная статистика столбцов Emotion, Information, Motivation представлена в табл. 3.

Для столбцов Emotion, Information, Motivation проверим попарную корреляцию Пирсона (r), которая показала следующие результаты: Emotion-Information $r = 0{,}072$; Emotion-Motivation $r = -0{,}031$; Information-Emotion $r = 0{,}038$ (для всех результатов $p < 0{,}05$). Исходя из полученных результатов корреляции Пирсона взаимосвязь между признаками не обнаружена.

Таким образом, на вход обучаемой модели поступают три значения — нормализованное число постов по каждому классу, на выход — значения {–1; 0; 1} соответствующие значению одного из факторов (A–Q4). В качестве используемых моделей выступают стандартные модели из библиотеки scikit-learn: LinearSVC (метод опорных векторов [30]), RandomForestClassifier

[1] Natural Language Toolkit [Электронный ресурс]. URL: https://www.nltk.org/ (дата обращения: 22.02.2023).

[2] Regular expression operations [Электронный ресурс]. URL: https://docs.python.org/3/library/re.html (дата обращения: 22.02.2023).

[3] Morphological analyzer pymorphy2 [Электронный ресурс]. URL: https://pymorphy2.readthedocs.io/en/stable/ (дата обращения: 22.02.2023).

[4] An open-source conversational AI framework DeepPavlov [Электронный ресурс]. URL: https://deeppavlov.ai/ (дата обращения: 22.02.2023).

[5] An open-source software library for machine learning and artificial intelligence [Электронный ресурс]. URL: https://www.tensorflow.org/ (дата обращения: 22.02.2023).

Таблица 1. Пример размеченных постов
Table 1. Example of marked up posts

Индекс строки (ind)	Идентификатор пользователя «ВКонтакте» (ID)	Подкласс поста		
		Эмоциональный (Emotion)	Информационный (Information)	Побудительно-деятельностный (Motivation)
0	10033***	Positive	Personal	Selling
1	10033***	No emotion	No information	No motivation
2	10033***	No emotion	Personal	No motivation
	…	…	…	…
27033	9917***	Positive	Personal	No motivation
27034	9917***	Negative	Personal	No motivation
27035	9917***	No emotion	Formal/statistical	No motivation

Таблица 2. Фрагмент финального набора данных по пользователям
Table 2. Fragment of the final user data set

ind	A	B	…	Q3	Q4	ID	fPost	dPost	Emotion	Information	Motivation
0	1	–1	…	–1	0	6524**	2716	2516	0,857	0,957	0,039
1	1	1	…	–1	0	1189**	1574	1382	0,840	0,942	0,047
2	1	0	…	0	0	4215**	1556	1369	0,793	0,953	0,085
3	1	–1	…	–1	1	4204**	1225	1099	0,830	0,978	0,009
4	1	1	…	0	0	4723**	975	899	0,812	0,972	0,039

Таблица 3. Первичные статистики столбцов
Table 3. Primary columns statistic

Статистика	Emotion	Information	Motivation
Минимум	0,227	0,321	0
25 перцентиль	0,650	0,517	0,081
50 перцентиль	0,751	0,628	0,179
75 перцентиль	0,889	0,986	0,356
Максимум	1	1	1

(случайный лес [31]) и MultinomialNB (наивный байесовский классификатор [32] в multi-class представлении).

В качестве метрик оценки классифицирующей модели использованы специфичные для задачи multi-class классификации метрики [33]: Accuracy, F1-micro и F1-macro. Данные метрики описываются таблицей сопряженности (табл. 4) для каждого из K классов {–1; 0; 1}. Для оценки качества классификации метрики высчитываются в формате 1vsALL для каждого класса. Например, для класса –1, классы 0 и 1 объединяются в класс 0. Таким образом, возможно использование таблицы сопряженности в бинарном виде.

В соответствии с табл. 1 введем метрику Accuracy:

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} \text{TP}_i + \text{TN}_i}{n_{\text{samples}}}, \qquad (1)$$

где K — количество классов (три); n_{samples} — размер выборки.

Таблица 4. Таблица сопряженности [31]
Table 4. Convergence table [31]

Категория значений		Предсказанное значение	
		Положительное	Негативное
Реальное значение	Положительное	TP	FN
	Негативное	FP	TN

Таблица 5. Результаты эксперимента для моделей машинного обучения LinearSVC/RandomForestClassifier/MultinomialNB, %
Table 5. Experimental results for machine learning models LinearSVC/RandomForestClassifier/MultinomialNB, %

Факторы	Метрики		
	Accuracy	F1-micro	F1-macro
A	0,69/**0,82**/0,51	0,60/**0,79**/0,45	0,48/**0,73**/0,39
B	0,66/**0,85**/0,54	0,60/**0,80**/0,53	0,38/**0,74**/0,38
C	0,44/0,74/0,38	0,44/0,74/0,38	0,20/0,74/0,23
E	0,44/0,78/0,32	0,44/0,78/0,32	0,25/0,78/0,31
F	0,49/**0,85**/0,47	0,44/**0,79**/0,44	0,33/**0,77**/0,31
G	0,39/0,71/0,38	0,39/0,71/0,38	0,20/0,70/0,30
H	0,53/0,77/0,43	0,53/0,77/0,43	0,23/0,72/0,26
I	0,61/**0,87**/0,41	0,59/**0,81**/0,38	0,25/**0,78**/0,36
L	0,40/0,73/0,38	0,40/0,73/0,38	0,25/0,73/0,27
M	0,54/0,76/0,56	0,54/0,76/0,56	0,23/0,72/0,30
N	0,49/**0,77**/0,36	0,47/**0,73**/0,35	0,34/**0,69**/0,31
O	0,47/0,76/0,36	0,47/0,76/0,36	0,21/0,75/0,34
Q1	0,47/**0,76**/0,26	0,43/**0,71**/0,24	0,31/**0,67**/0,23
Q2	0,42/0,73/0,35	0,42/0,73/0,35	0,28/0,73/0,34
Q3	0,45/0,71/0,43	0,45/0,71/0,43	0,26/0,68/0,34
Q4	0,40/0,73/0,39	0,40/0,73/0,39	0,23/0,72/0,27

Значение метрики Accuracy лежит в пределах от 0 до 1 (чем выше значение, тем лучше). Данная метрика предсказывает долю правильных ответов классификатора по всем классам вне зависимости от их сбалансированности [31].

Рассчитаем метрики F1-micro и F1-macro [33]:

$$\text{F1-micro} = \frac{\sum_{i=1}^{K} \text{TP}_i}{n_{\text{samples}}}, \quad (2)$$

$$\text{F1-macro} = 2\left(\frac{\text{MacroPrecision} \times \text{MacroRecall}}{\text{MacroPrecision}^{-1} + \text{MacroRecall}^{-1}}\right), \quad (3)$$

где

$$\text{MacroPrecision} = \frac{\sum_{i=1}^{K} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}}{K},$$

$$\text{MacroRecall} = \frac{\sum_{i=1}^{K} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}}{K}.$$

Метрика F1-micro (2) в случае отнесения моделью всех элементов выборки к одному классу будет равна метрике (1). Именно поэтому она позволяет определить разделяющую силу построенной модели. Метрика F1-macro (3) напротив позволяет оценить общее качество классификатора на всех данных с оценкой в промежутке [0; 1], где 0 — худшее значение метрики, а 1 — лучшее. Тестирование полученных классифицирующих моделей проведено с помощью метода скользящего контроля по четырем блокам. В соответствии со спецификацией данного метода [34] разделим выборку на четыре непересекающихся блока. Каждый блок по очереди является контрольной подвыборкой, при этом обучение производится по остальным трем блокам.

Вычислительный эксперимент

В табл. 5 представлены результаты, полученные при усреднении метрик, полученных на четырех контрольных подвыборках для каждого из факторов. В столбцах указаны факторы от A до Q4, в строках — значения метрики (Accuracy, F1-micro, F1-macro). Значения в ячейках показаны через косую черту для каждой модели машинного обучения, LinearSVC/RandomForestClassifier/MultinomialNB. Наилучшее значение оценки классифицирующей модели для наглядности выделено полужирным шрифтом. Заметим, что модель RandomForestClassifier показала наилучшее качество классификации. Для некоторых факторов (C, E, G, H, L, M, O, Q2, Q3, Q4), где отсутствует выделение полужирным шрифтом значений в ячейках, не удалось построить достоверную модель для предсказаний.

Обсуждение

Для уточнения предсказаний и верификации результатов необходимо существенно расширить (как минимум на один порядок) число респондентов, так как на данном этапе построения моделей, учитывающих подклассы главных классов (для информационных: формальные, событийные, личные, интеллектуально-рассудительные, ссылочные, кулинарные; эмоциональных: позитивные, негативные и поздравительные;

побудительно-деятельностных: благотворительные, продающие, побудительные к действию) не хватает данных для корректного обучения. Однако стоит отметить, что по результатам исследования [16] взаимосвязь между распределением постов и выраженностью некоторых факторов выявлена статистически.

Исходя из количества пользователей в начале эксперимента (312 человек) и в его конце (183 человека), можно сделать вывод, что предлагаемая автоматизация только по пользовательским постам может быть сильно ограничена в применимости. Возможностью для будущих исследований является разработка моделей, подходов и методов, которые будут опираться не только на посты, но и на другие доступные пользовательские данные.

Из-за использования в процессе автоматизации предсказания оценки выраженности психологических особенностей моделей для классификации постов, которые были представлены в статье авторов [19] и кратко описаны в начале раздела «Формирование и описание набора данных, метрики», может возникать смещение результатов из-за ошибок этих моделей. В дальнейшим данный вопрос требует изучения, в том числе при помощи сравнения с моделями, которые будут работать не при помощи классификации постов, а напрямую при помощи эмбеддингов, извлекаемых из них.

Результаты исследования продемонстрировали целесообразность в дальнейшем провести сравнение представленного в настоящей работе подхода с подходами на основе использования только комбинаций эмбеддингов постов, получаемых при помощи BERT, ELMO или родственных языковых моделей, минуя этап их классификации, а также с подходами на основе дистрибутивно-семантических языковых моделей [35]. Полученная модель может иметь ограничения применимости, связанные с существенно различающимся количеством постов у разных пользователей, которые будут изучены в дальнейших исследованиях. Возможно, данная модель будет лучше работать на аккаунтах пользователей, имеющих заметное число постов, превышающее определенный порог или начиная с заданного порога публикационной интенсивности. Полученные разработки планируется внедрить в прототипе комплекса для анализа пользователей социальных сетей[1].

Заключение

В работе представлены результаты построения предсказательной модели для оценки выраженности личностных особенностей пользователей в соответствии с тестом Р. Кеттелла на основе анализа, публикуемого пользователями в социальных сетях контента. Финальный набор данных для построения моделей и дальнейшего тестирования их работы составлен из 183 респондентов, прошедших тест Р. Кеттелла, со ссылками на их открытые аккаунты в социальной сети. Построенные модели могут предсказывать значение факторов A, B, F, I, N, Q1, а значения остальных факторов (C, E, G, H, L, M, O, Q2, Q3, Q4) модели не различают, относя все результаты к одному классу (на это указывает равенство результатов метрики F1-micro и Accuracy). Текущий набор данных явным образом недостаточен для построения качественных классификаторов по всем факторам 16-факторного теста Р. Кеттелла, таким образом можно заключить, что требуется увеличение исходного набора данных как минимум на один порядок. Теоретическая значимость заключается в разработке и проверке нового подхода (включающего в себя языковые модели и нейронные сети), который позволит автоматизировать процесс оценки выраженности личностных особенностей пользователей социально сети. Практическая значимость заключается в создании наработок для прототипа автоматизированной системы предсказания оценки выраженности психологических особенностей пользователей социальной сети. Дальнейшими направлениями исследования являются существенное расширение количества респондентов для увеличения числа предсказываемых факторов и дополнение моделей новыми признаками (со страниц пользователей в социальной сети) для повышения точности и применимости.

[1] Комплекс для анализа пользователей социальной сети [Электронный ресурс]. URL: https://sea.dscs.pro/ (дата обращения: 22.02.2023).

Литература

1. Vander Shee B.A., Peltier J., Dahl A.J. Antecedent consumer factors, consequential branding outcomes and measures of online consumer engagement: Current research and future directions // Journal of Research in Interactive Marketing. 2020. V. 14. N 2. P. 239–268. https://doi.org/10.1108/JRIM-01-2020-0010
2. Fayaz A., Muhammad Z.T., Ayaz A. The Big Five dyad congruence and compulsive buying: A case of service encounters // Journal of Retailing and Consumer Services. 2022. V. 68. P. 103007. https://doi.org/10.1016/j.jretconser.2022.103007
3. Shanahan T., Tran T.P., Taylor E.C. Getting to know you: Social media personalization as a means of enhancing brand loyalty and perceived quality // Journal of Retailing and Consumer Services. 2019. N 47. P. 57–65. https://doi.org/10.1016/j.jretconser.2018.10.007
4. Woods S.A., Mustafa M.J., Anderson N., Sayer B. Innovative work behavior and personality traits: Examining the moderating effects of organizational tenure // Journal of Managerial Psychology. 2018. V. 33. N 1. P. 29–42. https://doi.org/10.1108/JMP-01-2017-0016
5. Bouiri O., Lotfi S., Talbi M. Correlative study between personality traits, student mental skills and educational outcomes // Education

References

1. Vander Shee B.A., Peltier J., Dahl A.J. Antecedent consumer factors, consequential branding outcomes and measures of online consumer engagement: Current research and future directions. *Journal of Research in Interactive Marketing*, 2020, vol. 14, no. 2, pp. 239–268. https://doi.org/10.1108/JRIM-01-2020-0010
2. Fayaz A., Muhammad Z.T., Ayaz A. The Big Five dyad congruence and compulsive buying: A case of service encounters. *Journal of Retailing and Consumer Services*, 2022, vol. 68, pp. 103007. https://doi.org/10.1016/j.jretconser.2022.103007
3. Shanahan T., Tran T.P., Taylor E.C. Getting to know you: Social media personalization as a means of enhancing brand loyalty and perceived quality. *Journal of Retailing and Consumer Services*, 2019, no. 47, pp. 57–65. https://doi.org/10.1016/j.jretconser.2018.10.007
4. Woods S.A., Mustafa M.J., Anderson N., Sayer B. Innovative work behavior and personality traits: Examining the moderating effects of organizational tenure. *Journal of Managerial Psychology*, 2018, vol. 33, no. 1, pp. 29–42. https://doi.org/10.1108/JMP-01-2017-0016
5. Bouiri O., Lotfi S., Talbi M. Correlative study between personality traits, student mental skills and educational outcomes. *Education*

Sciences. 2021. V. 11. N 4. P. 153. https://doi.org/10.3390/educsci11040153
6. Chekalev A.A., Khlobystova A.O., Tulupyeva T.V. Applicant's decision support system for choosing the direction of study // Proc. of the XXV International Conference on Soft Computing and Measurements (SCM). 2022. P. 226–228. https://doi.org/10.1109/SCM55405.2022.9794902
7. Stoliarova V.F., Tulupyev A.L. Cumulative mean function of public posting episodes in the online media with regard to user's digital traces: Limited data on publications dates and profile data // Proc. of the XXV International Conference on Soft Computing and Measurements (SCM). 2022. P. 25–27. https://doi.org/10.1109/SCM55405.2022.9794894
8. Thielmann I., Spadaro G., Balliet D. Personality and prosocial behavior: A theoretical framework and meta-analysis // Psychological Bulletin. 2020. V. 146. N 1. P. 30–90. https://doi.org/10.1037/bul0000217
9. Clark C., Davila A., Regis M., Kraus S. Predictors of COVID-19 voluntary compliance behaviors: An international investigation // Global Transitions. 2020. V. 2. P. 76–82. https://doi.org/10.1016/j.glt.2020.06.003
10. Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users // Studies in Computational Intelligence. 2020. V. 868. P. 272–277. https://doi.org/10.1007/978-3-030-32258-8_32
11. Piotrowski C., Sherry D., Keller J.W. Psychodiagnostic test usage: A survey of the society for personality assessment // Journal of Personality Assessment. 1985. V. 49. N 2. P. 115–119. https://doi.org/10.1207/s15327752jpa4902_1
12. Goldber L.R. An alternative "description of personality": the big-five factor structure // Journal of Personality and Social Psychology. 1990. V. 59. N 6. P. 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216
13. Schwartz S.H. A proposal for measuring value orientations across nations // Questionnaire Development Package of the European Social Survey. 2003. N 259(290). P. 261–319.
14. Cattell H.E.P., Mead A.D. The sixteen personality factor questionnaire (16PF) // The SAGE Handbook of Personality Theory and Assessment. V. 2. 2008. P. 135–159. https://doi.org/10.4135/9781849200479.n7
15. Plutchik R., Kellerman H., Conte H.R. A structural theory of ego defenses and emotions // Emotions, Personality, and Psychotherapy. Boston: Springer, 1979. P. 227–257. https://doi.org/10.1007/978-1-4613-2892-6_9
16. Тулупьева Т.В., Тафинцева А.С., Тулупьев А.Л. Подход к анализу отражения особенностей личности в цифровых следах // Вестник психотерапии. 2016. № 60(65). С. 124–137.
17. Azucar D., Marengo D., Settanni M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis // Personality and Individual Differences. 2018. V. 124. P. 150–159. https://doi.org/10.1016/j.paid.2017.12.018
18. Oliseenko V.D., Tulupyeva T.V. Neural network approach in the task of multi-label classification of user posts in online social networks // Proc. of the XXIV International Conference on Soft Computing and Measurements (SCM). 2021. P. 46–48. https://doi.org/10.1109/SCM52931.2021.9507148
19. Oliseenko V.D., Eirich M., Tulupyev A.L., Tulupyeva T.V. BERT and ELMo in task of classifying social media users posts // Lecture Notes in Networks and Systems. 2023. V. 566. P. 475–486. https://doi.org/10.1007/978-3-031-19620-1_45
20. Tay L., Woo S.E., Hickman L., Saef R.M. Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining // European Journal of Personality. 2020. V. 34. N 5. P. 826–844. https://doi.org/10.1002/per.2290
21. Bleidorn W., Hopwood Ch.J. Using machine learning to advance personality assessment and theory // Personality and Social Psychology Review. 2019. V. 23. N 2. P. 190–203. https://doi.org/10.1177/1088868318772990
22. Kahn J.H., Tobin R.M., Massey A.E., Anderson J.A. Measuring emotional expression with the Linguistic Inquiry and Word Count // The American Journal of Psychology. 2007. V. 120. N 2. P. 263–286. https://doi.org/10.2307/20445398
23. Hartmann J., Huppertz J., Schamp C., Heitmann M. Comparing automated text classification methods // International Journal of

Sciences, 2021, vol. 11, no. 4, pp. 153. https://doi.org/10.3390/educsci11040153
6. Chekalev A.A., Khlobystova A.O., Tulupyeva T.V. Applicant's decision support system for choosing the direction of study. *Proc. of the XXV International Conference on Soft Computing and Measurements (SCM)*, 2022, pp. 226–228. https://doi.org/10.1109/SCM55405.2022.9794902
7. Stoliarova V.F., Tulupyev A.L. Cumulative mean function of public posting episodes in the online media with regard to user's digital traces: Limited data on publications dates and profile data. *Proc. of the XXV International Conference on Soft Computing and Measurements (SCM)*, 2022, pp. 25–27. https://doi.org/10.1109/SCM55405.2022.9794894
8. Thielmann I., Spadaro G., Balliet D. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, 2020, vol. 146, no. 1, pp. 30–90. https://doi.org/10.1037/bul0000217
9. Clark C., Davila A., Regis M., Kraus S. Predictors of COVID-19 voluntary compliance behaviors: An international investigation. *Global Transitions*, 2020, vol. 2, pp. 76–82. https://doi.org/10.1016/j.glt.2020.06.003
10. Khlobystova A.O., Abramov M.V., Tulupyev A.L. Soft estimates for social engineering attack propagation probabilities depending on interaction rates among instagram users. *Studies in Computational Intelligence*, 2020, vol. 868, pp. 272–277. https://doi.org/10.1007/978-3-030-32258-8_32
11. Piotrowski C., Sherry D., Keller J.W. Psychodiagnostic test usage: A survey of the society for personality assessment. *Journal of Personality Assessment*, 1985, vol. 49, no. 2, pp. 115–119. https://doi.org/10.1207/s15327752jpa4902_1
12. Goldber L.R. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, 1990, vol. 59, no. 6, pp. 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216
13. Schwartz S.H. A proposal for measuring value orientations across nations. *Questionnaire Development Package of the European Social Survey*, 2003, no. 259(290), pp. 261–319.
14. Cattell H.E.P., Mead A.D. The sixteen personality factor questionnaire (16PF). *The SAGE Handbook of Personality Theory and Assessment. V. 2*, 2008, pp. 135–159. https://doi.org/10.4135/9781849200479.n7
15. Plutchik R., Kellerman H., Conte H.R. A structural theory of ego defenses and emotions. *Emotions, Personality, and Psychotherapy*. Boston, Springer, 1979, pp. 227–257. https://doi.org/10.1007/978-1-4613-2892-6_9
16. Tulupyeva T.V., Tafintseva A.S., Tulupyev A.L. An approach to the analysis of personal traits reflection in digital traces. *Bulletin of Psychotherapy*, 2016, no. 60(65), pp. 124–137. (in Russian)
17. Azucar D., Marengo D., Settanni M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 2018, vol. 124, pp. 150–159. https://doi.org/10.1016/j.paid.2017.12.018
18. Oliseenko V.D., Tulupyeva T.V. Neural network approach in the task of multi-label classification of user posts in online social networks. *Proc. of the XXIV International Conference on Soft Computing and Measurements (SCM)*, 2021, pp. 46–48. https://doi.org/10.1109/SCM52931.2021.9507148
19. Oliseenko V.D., Eirich M., Tulupyev A.L., Tulupyeva T.V. BERT and ELMo in task of classifying social media users posts. *Lecture Notes in Networks and Systems*, 2023, vol. 566, pp. 475–486. https://doi.org/10.1007/978-3-031-19620-1_45
20. Tay L., Woo S.E., Hickman L., Saef R.M. Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, 2020, vol. 34, no. 5, pp. 826–844. https://doi.org/10.1002/per.2290
21. Bleidorn W., Hopwood Ch.J. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 2019, vol. 23, no. 2, pp. 190–203. https://doi.org/10.1177/1088868318772990
22. Kahn J.H., Tobin R.M., Massey A.E., Anderson J.A. Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, 2007, vol. 120, no. 2, pp. 263–286. https://doi.org/10.2307/20445398
23. Hartmann J., Huppertz J., Schamp C., Heitmann M. Comparing automated text classification methods. *International Journal of Research in Marketing*, 2019, vol. 36, no. 1, pp. 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

Research in Marketing. 2019. V. 36. N 1. P. 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009
24. Eichstaedt J.C., Kern M.L., Yaden D.B., Schwartz H.A., Giorgi S., Park G., Hagan C.A., Tobolsky V.A., Smith L.K., Buffone A., Iwry J., Seligman M.E.P., Ungar L.H. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations // Psychological Methods. 2021. V. 26. N 4. P. 398–427. https://doi.org/10.1037/met0000349
25. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). V. 1. 2019. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423
26. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations // Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). V. 1. 2018. P. 2227–2237. https://doi.org/10.18653/v1/N18-1202
27. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner Ch., McCandlish S., Radford A., Sutskever I., Amodei D. Language models are few-shot learners // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 2020.
28. Sun J., Tian Z., Fu Y., Geng J., Liu C. Digital twins in human understanding: a deep learning-based method to recognize personality traits // International Journal of Computer Integrated Manufacturing. 2021. V. 34. N 7-8. P. 860–873. https://doi.org/10.1080/0951192X.2020.1757155
29. Wang Z., Wu C.-H., Li Q.-B., Yan B., Zheng K.-F. Encoding text information with graph convolutional networks for personality recognition // Applied Science. 2020. V. 10. N 12. P. 4081. https://doi.org/10.3390/app10124081
30. Cortes C., Vapnik V. Support-vector networks // Machine Learning. 1995. V. 20. N 3. P. 273–297. https://doi.org/10.1023/A:1022627411411
31. Breiman L. Random forests // Machine Learning. 2001. V. 45. N 1. P. 5–32. https://doi.org/10.1023/A:1010933404324
32. Friedman N., Geiger D., Goldszmidt M. Bayesian network classifiers // Machine Learning. 1997. V. 29. N 2-3. P. 131–163. https://doi.org/10.1023/a:1007465528199
33. Grandini M., Bagli E., Visani G. Metrics for multi-class classification: an overview. 2020 [Электронный ресурс]. URL: https://arxiv.org/abs/2008.05756 (дата обращения: 01.09.2022).
34. Refaeilzadeh P., Tang L., Liu H. Cross-Validation // Encyclopedia of Database Systems. Boston: Springer, 2009. P. 532–538. https://doi.org/10.1007/978-0-387-39940-9_565
35. Груздева А.С., Бессмертный И.А. Классификация коротких текстов с использованием волновой модели // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22. № 2. С. 287–293. https://doi.org/10.17586/2226-1494-2022-22-2-287-293

24. Eichstaedt J.C., Kern M.L., Yaden D.B., Schwartz H.A., Giorgi S., Park G., Hagan C.A., Tobolsky V.A., Smith L.K., Buffone A., Iwry J., Seligman M.E.P., Ungar L.H. Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 2021, vol. 26, no. 4, pp. 398–427. https://doi.org/10.1037/met0000349
25. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). V. 1*, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
26. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations. *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). V. 1*, 2018, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202
27. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner Ch., McCandlish S., Radford A., Sutskever I., Amodei D. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.
28. Sun J., Tian Z., Fu Y., Geng J., Liu C. Digital twins in human understanding: a deep learning-based method to recognize personality traits. *International Journal of Computer Integrated Manufacturing*, 2021, vol. 34, no. 7-8, pp. 860–873. https://doi.org/10.1080/0951192X.2020.1757155
29. Wang Z., Wu C.-H., Li Q.-B., Yan B., Zheng K.-F. Encoding text information with graph convolutional networks for personality recognition. *Applied Science*, 2020, vol. 10, no. 12, pp. 4081. https://doi.org/10.3390/app10124081
30. Cortes C., Vapnik V. Support-vector networks. *Machine Learning*, 1995, vol. 20, no. 3, pp. 273–297. https://doi.org/10.1023/A:1022627411411
31. Breiman L. Random forests. *Machine Learning*, 2001, vol. 45, no. 1, pp. 5–32. https://doi.org/10.1023/A:1010933404324
32. Friedman N., Geiger D., Goldszmidt M. Bayesian network classifiers. *Machine Learning*, 1997, vol. 29, no. 2-3, pp. 131–163. https://doi.org/10.1023/a:1007465528199
33. Grandini M., Bagli E., Visani G. *Metrics for multi-class classification: an overview*. 2020. Available at: https://arxiv.org/abs/2008.05756 (accessed: 01.09.2022).
34. Refaeilzadeh P., Tang L., Liu H. Cross-Validation. *Encyclopedia of Database Systems*. Boston, Springer, 2009, pp. 532–538. https://doi.org/10.1007/978-0-387-39940-9_565
35. Gruzdeva A.S., Bessmertny I.A. Classification of short texts using a wave model. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2022, vol. 22, no. 2, pp. 287–293. (in Russian). https://doi.org/10.17586/2226-1494-2022-22-2-287-293

Авторы

Олисеенко Валерий Дмитриевич — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57219554703, https://orcid.org/0000-0002-3479-0085, vdo@dscs.pro

Абрамов Максим Викторович — кандидат технических наук, руководитель лаборатории, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 56938320500, https://orcid.org/0000-0002-5476-3025, mva@dscs.pro

Authors

Valerii D. Oliseenko — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 57219554703, https://orcid.org/0000-0002-3479-0085, vdo@dscs.pro

Maxim V. Abramov — PhD, Head of Laboratory, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 56938320500, https://orcid.org/0000-0002-5476-3025, mva@dscs.pro

Статья поступила в редакцию 29.09.2022
Одобрена после рецензирования 08.12.2022
Принята к печати 17.03.2023

Received 29.09.2022
Approved after reviewing 08.12.2022
Accepted 17.03.2023