

Анализ соответствия сообществ пользователя его профессиональному профилю на основе методов обработки естественного языка

А. О. Иващенко

Санкт-Петербургский
Федеральный исследовательский
центр Российской академии наук
aok@dscs.pro

Е. А. Годзун

Санкт-Петербургский
государственный университет
egorgodzun@gmail.com

М. В. Абрамов

Санкт-Петербургский
Федеральный исследовательский
центр Российской академии наук
mva@dscs.pro

Аннотация. Целью работы является разработка модели, определяющей социально-профессиональные предпочтения пользователей социальной сети «ВКонтакте» на основе их подписок на сообщества. Предлагаемый подход основывается на сопоставлении текстовых описаний сообществ с шестью типами личности по Дж. Голланду на основе методов обработки естественного языка. Экспериментальные результаты, полученные на выборке из 965 пользователей, показали высокую эффективность подхода: в 95,9 % случаев хотя бы один из трёх наиболее выраженных типов, определённых моделью, совпадает с итогами классического теста Голланда. Предложенное решение способно упростить процесс первичной профориентации, предоставляя релевантные рекомендации на основе цифровых следов пользователей.

Ключевые слова: искусственный интеллект, обработка естественного языка, профориентация, модель Голланда, социальные сети, цифровые следы

I. ВВЕДЕНИЕ

Профессиональное самоопределение является актуальной проблемой как молодых граждан – выпускников школ и вузов, так и для зрелых людей, рассматривающих смену карьеры. По данным Росстата, у 23,8 % выпускников работа не связана с полученным высшим образованием [1]. Это может свидетельствовать как о недостаточной эффективности системы профориентации, так и о сложности выбора профессии на этапе поступления в вуз. Среди взрослого населения ситуация также вызывает беспокойство: среди 1790 опрошенных россиян в возрасте от 50 до 67 лет доля, задумывающихся о кардинальной смене профессии составляет 53 % [2].

Традиционные методы профориентации включают консультации специалистов и применение специализированных психологических тестов, позволяющих оценить личностные склонности и профессиональные предпочтения [3]. **Ошибка! Источник ссылки не найден.** В этом процессе могут помочь консультации психологов или прохождение различных профориентационных тестов. Однако эти методы обладают определенными недостатками, среди

которых необходимость высокой вовлечённости в процесс самого человека, и как следствие, существенные затраты личного времени.

В связи с этим возрастает интерес к автоматизированным методам профориентации, опирающимся на анализ цифровых данных, в частности, данных онлайн-социальных медиа. В последние годы социальные сети стали не только средством коммуникации, но и мощным инструментом анализа профессиональной деятельности и интересов пользователей. Так социальные сети позволяют рекрутерам оценивать профессиональную репутацию и активность кандидатов, анализируя их участие в профессиональных группах, обсуждениях и публикациях [4].

Одним из подходов к анализу профессиональных склонностей является использование модели RIASEC, предложенной Дж. Голландом, которая классифицирует личности по шести основным типам профессиональной направленности: реалистический, исследовательский, артистический, социальный, предприимчивый и конвенциональный [5]. Состоятельность такого подхода описана в исследованиях [6], [7].

В данной работе предлагается метод автоматизированного определения социально-профессиональной направленности личности по модели RIASEC на основе анализа подписок пользователя в онлайн социальных медиа. В основе исследования лежит гипотеза о том, что тематики сообществ, на которые подписан пользователь, могут являться индикатором его профессиональных интересов и склонностей [8]. Основной целью исследования является разработка модели, определяющей социально-профессиональные предпочтения пользователей социальной сети «ВКонтакте» на основе их подписок на сообщества.

II. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Современные исследования подтверждают, что цифровые следы пользователей в социальных сетях могут использоваться для анализа их личности и профессиональных предпочтений. В работе [9] было продемонстрировано, что машинное обучение, использующее анкетные данные, позволяет с высокой точностью предсказывать профессиональную склонность пользователей. В исследовании [10]

анализировались записи 2834 пользователей с применением лингвистического словаря LIWC и тематического моделирования LDA. Результаты показали умеренную точность предсказания профессиональных интересов, но подтвердили жизнеспособность NLP-методов. В исследовании [11] с помощью нейросетевых моделей (LSTM, GRU) классифицировались посты по профессиональной сфере, достигнув точности 93–94 %. Также существует ряд работ по прогнозированию личности на основе цифровых следов пользователя в социальных сетях с использованием алгоритмов машинного обучения [12], [13], [14]. В работе [15] авторы приходят к выводу, что сочетание анализа тем и эмоционального контекста с LSTM и механизмом внимания позволяет значительно повысить точность предсказания личности. В работе [16] рассматривается возможность построения оценок знаний, навыков и профессиональных интересов на основе текстовых описаний профессий с применением NLP, что демонстрирует перспективность использования подобных методов в задачах профориентации.

Настоящее исследование продолжает работу [8], где было выявлено, что интересы пользователей, выраженные через тематики сообщений, на которые они подписаны в социальной сети «ВКонтакте», имеют связь с результатами теста Голланда. Основной вклад данной работы заключается в разработке модели предсказания типов RIASEC на основе анализа текстового описания подписок пользователей, что позволяет сопоставлять интересы пользователей с типами профессий. Этот подход открывает новые возможности для автоматизированной профориентации, более точного карьерного консультирования и адаптации профессионального подбора специалистов на основе их цифровых следов.

III. МАТЕРИАЛЫ И МЕТОДЫ

В основе исследования лежат результаты тестирования по модели Дж. Голланда по определению выраженности у человека шести типов личности: реалистический (R), исследовательский (I), артистический (A), социальный (S), предприимчивый (E) и конвенциональный (C). Каждому типу соответствует род деятельности, к которой человек данного типа имеет наибольшую предрасположенность и способности [5]. Тест состоит из 42 пар профессий. Из каждой пары необходимо выбрать одну профессию. Результаты прохождения теста являются значениями для типов RIASEC в диапазоне от 1 до 14 (рис. 1). По результатам теста выделяется так называемая «ведущая триада» – три типа личности с наибольшими значениями баллов, которые наиболее полно отражают склонность индивида к определённым видам профессиональной деятельности.

A. Описание набора данных

Данные были собраны при помощи опроса, размещенного в веб-приложении на базе сервиса VK Mini Apps – «Психологические тесты» (<https://vk.com/app7794698>). Приложение было доступно всем пользователям онлайн социальной сети «ВКонтакте». Исследование проводилось на данных, сформированных из результатов 965 пользователей «ВКонтакте», которые прошли тест Голланда и дали

согласие на обработку полученных результатов. Пример рассматриваемого набора данных представлен в табл. I, где представлены идентификатор пользователя (id), пол, количество друзей, типы личности (R, I, A, S, E, C), значения которых были получены в результате прохождения теста, данные о группах, куда входят идентификатор группы (group_id), ее название (name), псевдоним (screen_name), закрыта группа или нет (is_closed), количество подписчиков (members_count), описание (description), статус (status), категория (activity), тип (type).

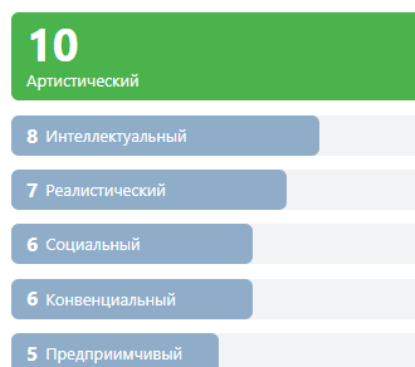


Рис. 1. Пример результата теста Голланда

Перед обработкой данных были удалены пользователи с закрытыми аккаунтами и пользователи, у которых количество подписок на группы меньше 10. Также для каждого пользователя были удалены группы, в которых отсутствовало описание и проведено укрупнение категорий групп на основе имеющейся категоризации «ВКонтакте».

ТАБЛИЦА I. ФРАГМЕНТ РАССМАТРИВАЕМОГО НАБОРА ДАННЫХ

id	Пол	Кол-во друзей	R	I	A	S	E	C	Данные о группах
1	2	29	8	5	9	8	7	5	[{"group_id": 180721685, "name": "DSCS.pro", "screen_name": "dscspro", "activity": "Образование", "description": "DSCS.pro производит комплексные высокотехноло...}]

B. Описание реализованной модели

Для проведения исследования использовался язык Python версии 3.11.8. Для предобработки текста использовались следующие библиотеки:

1. re – библиотека для работы с регулярными выражениями. С ее помощью из текстов были удалены ссылки, почты, смайлики, хештеги, цифры и др.;

2. spacy – библиотека для расширенной обработки естественного языка в Python. Из spacy было взято множество стоп-слов для русского языка;

3. natasha – библиотека, которая решает основные задачи NLP для русского языка. С ее помощью производилась лемматизация текста.

4. PyTorch – фреймворк машинного обучения для языка Python с открытым исходным кодом. С помощью PyTorch – токенизация и получение эмбеддингов.

В данной исследовательской работе для получения эмбеддингов использовались RoBERTa (Robustly optimized BERT approach), являющейся улучшенной версией BERT, а также сам BERT.

Для решения задачи сопоставления групп, на которые подписан пользователь, и типа RIASEC было сделано следующее. Для каждой группы были взяты её описание, статус, название и категория. Типы RIASEC были представлены через их текстовые описания, взятые из «Опросника профессиональных предпочтений по Голланду» [17]. Затем была проведена предобработка и лемматизация данных с использованием моделей ruRoBERTa-large и ruBERT-base. Для определения соответствия между группой и типом RIASEC применялось вычисление косинусного расстояния между векторами признаков. После производилась оценка модели по следующей метрике: определялись ведущие триады предсказанных и фактических значений для каждого участника эксперимента, и подсчитывалось количество их пересечений (рис. 2).

Полная схема модели представлена на рис. 3.



Рис. 2. Пример подсчета пересечений между предсказанными и истинными ведущими триадами

IV. РЕЗУЛЬТАТЫ

В ходе сравнения моделей ruRoBERTa-large и ruBERT-base по предложенной метрике (рис. 2) лучшие результаты показала ruBERT-base. Из результатов, полученных при использовании модели ruBERT-base

(табл. II), следует, что наличие хотя бы одного пересечения между ведущими триадами наблюдалось в 95,9 % случаях, в том числе для двойного или тройного пересечения – 52,6 %.

ТАБЛИЦА II. КОЛИЧЕСТВО ПЕРЕСЕЧЕНИЙ ВЕДУЩИХ ТРИАД ТИПОВ RIASEC ДЛЯ ПРЕДСКАЗАННЫХ ФАКТИЧЕСКИХ ЗНАЧЕНИЙ

Количество совпадающих типов	Количество правильно предсказанных значений	
	ruBERT-base	ruRoBERTa-large
2	319	313
1	293	315
3	37	23
0	28	26

Также в ходе эксперимента было вычислено количество уникальных категорий, которые попали в каждый тип, и общее количество групп для каждого типа (табл. III). Категории «Развлечения», «Увлечения и хобби», «Персона», «Красота, здоровье», «Образование» присутствуют у каждого типа, «Телекоммуникации» и «Высокие технологии» только у исследовательского (I), «Инвестиции» и «Розничная торговля» – предпринимчивого (E) типа, «Звукопись» и «Фото- и видеосъемка» – артистического (A) типа. Преобладание конвенционального типа (C) может указывать на то, что в данной выборке большинство пользователей связаны с офисной работой. Высокая представленность артистического типа (A) может свидетельствовать о значительном интересе пользователей к творческим профессиям, таким как дизайн, музыка, литература и искусство.

ТАБЛИЦА III. КОЛИЧЕСТВО КАТЕГОРИЙ И ГРУПП ДЛЯ ТИПОВ RIASEC

Тип	Количество категорий	Общее количество групп
C	70	42260
A	63	23422
I	52	17843
E	57	14770
R	46	4570
S	38	3616

Полученные результаты говорят о том, что модель, основанная на подписках пользователя, быть использована для предсказания трёх ведущих типов социально-профессиональной направленности. Высокий процент совпадений подтверждает, что такие рекомендации будут релевантными и полезными для пользователя. Вместе с тем, стоит отметить, данное решение может быть использовано для первичной диагностики, для получения точных результатов необходима консультация с психологом.

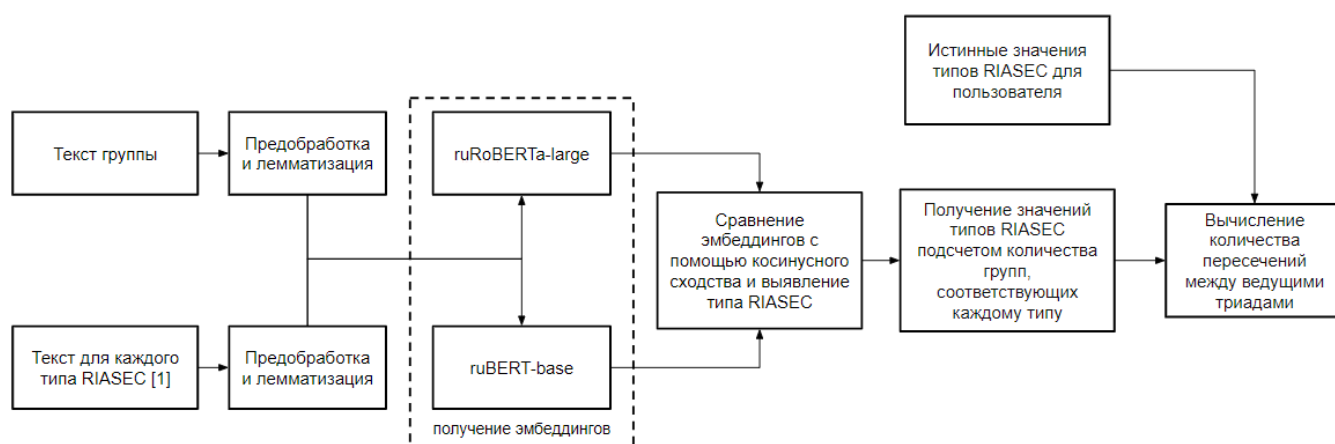


Рис. 3. Схема разработанной модели

V. ЗАКЛЮЧЕНИЕ

В данном исследовании был предложен подход для определения типа социально-профессиональной направленности, в частности разработана модель для автоматизированного определения социально-профессиональных предпочтений пользователей «ВКонтакте» на основе их цифровых следов, таких как подписки на группы. В основе модели лежат современные методы обработки естественного языка.

Результаты показали, что модель успешно определяет соответствие интересов пользователей типам личности по тесту Голланда (RIASEC). Совпадения хотя бы одного из предсказанных типов с фактическими были обнаружены в 95,9 % случаев, что свидетельствует о высокой точности рекомендаций. Значимость результата заключается в том, что предложенная модель может существенно помочь пользователям в осознанном выборе профессии, предоставляя рекомендации, основанные на анализе их цифровых следов.

В перспективах исследования планируется улучшение предсказательной способности модели за счет расширения анализируемых цифровых следов.

СПИСОК ЛИТЕРАТУРЫ

- [1] Рабочая сила, занятость и безработица в России 2024 [Электронный ресурс]. Режим доступа: <https://rosstat.gov.ru/folder/210/document/13211>, свободный. Яз. Рус. (дата обращения 15.05.2024);
- [2] Skillfactory и «Зарплата.ру» опросили россиян старше 50 лет о желании сменить работы на IT или другую отрасль [Электронный ресурс]. Режим доступа: <https://habr.com/ru/news/794793/>, свободный. Яз. Англ. (дата обращения 15.11.2024)
- [3] Прохоров А.В. Современные подходы к профессиональной ориентации школьников // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2022. Т. 27. № 2. С. 319–328.
- [4] Ruparel N., Dhir A., Tandon A., Kaur P., Islam J.U. The influence of online professional social media in human resource management: A systematic literature review // Technology in Society. 2020. Vol. 63. P. 101335. DOI: 10.1016/j.techsoc.2020.101335.
- [5] Holland J. The Psychology of Vocational Choice: A Theory of Personality Types and Model Environments. Blaisdell Publishing Company, 1966. 132 p.
- [6] Sheldon K.M., Holliday G., Titova L., Benson C. Comparing Holland and Self-Determination Theory measures of career preference as

predictors of career choice // Journal of Career Assessment. 2020. Vol. 28. № 1. Pp. 28–42. DOI: 10.1177/1069072718823003.

- [7] Batista J., Guedes-Gondim S. Personality and Person-Work Environment Fit: A Study Based on the RIASEC Model // International Journal of Environmental Research and Public Health. 2022. Vol. 20. № 1. DOI: 10.3390/ijerph20010719.
- [8] Хлобыстова А.О., Абрамов М.В., Столярова В.Ф. Исследование тенденций взаимосвязи между профориентационными предпочтениями пользователей и их цифровыми следами в социальной сети // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 3. С. 564–574. doi: 10.17586/2226-1494-2023-23-3-564-574
- [9] Kiselev P., Kiselev B., Matsuta V., Feshchenko A., Bogdanovskaya I., Kosheleva A. Career guidance based on machine learning: social networks in professional identity construction // Procedia Computer Science, 2020. Vol. 169. Pp. 158–163. DOI: 10.1016/j.procs.2020.02.128
- [10] Du Y.Y.L., Jain D., Cho Y.M., Hou D.X., Guntuku S.C., Ungar L., Tay L. Exploring a language-based interest assessment: Predicting vocational interests on social media using natural language processing // Journal of Career Assessment. 2024. P. 10690727241289125. DOI: 10.1177/1069072724128912
- [11] Dağıstanlı Ö., Erbay H., Kör H., Yurttakal A.H. Reflection of people's professions on social media platforms //Neural Computing and Applications. 2023. Vol. 35. №. 7. Pp. 5575-5586.
- [12] Oliseenko V.D., Khlobystova A.O., Korepanova A.A., Tulupyeva T.V. Automating the temperament assessment of online social network users // Doklady Mathematics. Moscow Pleiades Publishing. 2024. Pp. 1–6. DOI: 10.1134/S1064562423701041
- [13] Iwasaki T., Seki Y., Kashino W., Keyaki A., Kando N. Estimating Citizen Personality Traits Using Social Media Posts // International Conference on Asian Digital Libraries, Springer, Singapore. 2025. Pp. 119–135. DOI: 10.1007/978-981-96-0868-3_10.
- [14] Kumar R.P., B. Mohan G.B., Sai G.D. Ensemble machine learning models in predicting personality traits and insights using Myers-Briggs dataset // 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2023. Pp. 1–7. DOI: 10.1109/ACCAI58221.2023.10199294.
- [15] Zhao J., Zeng D., Xiao Y., Che L., Wang M. User personality prediction based on topic preference and sentiment analysis using LSTM model //Pattern Recognition Letters. 2020. Vol. 138. Pp. 397–402. DOI: 10.1016/j.patrec.2020.07.035
- [16] Putka D.J., Oswald F.L., Landers R.N., Beatty A.S., McCloy R.A., Yu M.C. Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings // Journal of Business and Psychology. 2023. Vol. 38. №. 2. Pp. 385–410. DOI: 10.1007/s10869-022-09824-0
- [17] Воробьев А.Н. Опросник профессиональных предпочтений по Холланду: Учебно-методическое пособие. М.: Когнитив-Центр, 2001, 20 с.