ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.82 *Научная статья* DOI: 10.18287/2223-9537-2022-12-1-82-92

Алгоритм психолингвистического анализа текстовых данных социальных сетей с применением модели «Большая пятёрка»

© 2022, Н.Г. Ярушкина, В.С. Мошкин ✉, И.А. Андреев

Ульяновский государственный технический университет, Ульяновск, Россия

Аннотация

Представлен подход к определению психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений в социальных сетях. Предложенный в работе подход заключается в классификации авторских текстов пользователя с использованием машинного обучения. В качестве обучающих данных используются результаты анализа опросов пользователей в соответствии с моделью «Большая пятёрка», а также набор авторских текстовых данных со страниц социальных сетей. Опросник содержит парные высказывания, опрашиваемый определяет степень собственного согласия с тем или иным высказыванием по шкале от 0 до 4. К текстовым ресурсам, используемым в качестве входных данных для классификатора, были применены методы обработки текстов на естественном языке (*NLP*), а также задействована лингвистическая онтология *RuWordNet*, с целью нивелирования ряда особенностей текстов социальных сетей, например, наличие грамматических ошибок и эмотиконов, затрудняющих процесс семантического анализа. В качестве классификаторов использовались две модели: метод опорных векторов и метод случайного леса. Для оценки эффективности использовалась метрика площади под кривой ошибок (*AUC ROC*). В экспериментах использовались открытые текстовые данные более 1000 пользователей социальной сети.

Ключевые слова: *модель «Большая пятёрка», машинное обучение, социальные сети, психолингвистический анализ.*

Цитирование: *Ярушкина Н.Г., Мошкин В.С., Андреев И.А.* Алгоритм психолингвистического анализа текстовых данных социальных сетей с применением модели «Большая пятёрка». *Онтология проектирования*. 2022. Т.12, №1(43). С. 82-92. DOI: 10.18287/2223-9537-2022-12-1-82-92.

Финансирование: Работа выполнена при финансовой поддержке Минобрнауки России в рамках проекта №075-00233-20-05 от 03.11.2020 «Исследование интеллектуального предиктивного мульти-модального анализа больших данных и извлечения знаний из различных источников».

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Работа с социальными сетями может принести пользу при реализации функции системы управления персоналом компании, так как зачастую из социальных сетей о профессиональных и личностных качествах кандидата на конкретную должность можно узнать больше, чем из его резюме. В настоящее время сбор и/или содержательный анализ собранной в социальных сетях информации проводится вручную специалистами кадровых служб, что требует больших затрат времени и ограничивает объём обрабатываемой информации.

В работе [1] отмечается важность сбора информации о пользователях социальных сетей с целью предиктивного анализа и выявления психических расстройств, в дальнейшем выражающихся в самоповреждающем поведении, а также нарушении эмоциональной саморегуляции. Анализ социальной структуры интернет-аудитории в зависимости от поведения поль-

зователей в интернет-пространстве рассматривается в исследовании [2]. В работе [3] пользователи социальных сетей объединяются в аутентичную субкультуру, при этом личностные черты каждого пользователя определяются путём анализа открытых структурированных (анкета) и неструктурированных (сообщения, статусы) данных страницы. Личностные черты могут выступать в качестве предикторов и коррелянтов различных психических отклонений, поэтому их определение может использоваться как часть диагностики личностных и психических расстройств [4, 5].

В психологии «Большая пятёрка» – это пятифакторная модель личности, разработанная таким образом, чтобы из набора входящих в неё черт можно было составить структурированный портрет личности. Эта модель включает пять основных факторов, каждый из которых, в свою очередь, объединяет группу черт. Измерение показателей модели «Большая пятёрка» осуществляется при помощи теста с соответствующим названием – пятифакторный опросник личности. Пятифакторная модель позволяет описать личность структурированно, охватив её разные стороны. «Большая пятёрка» включает следующие основные факторы:

- нейротизм;
- экстраверсия;
- открытость опыту;
- согласие, или сотрудничество;
- сознательность, или добросовестность.

Доработанный вариант теста *5PFQ* представлен японским психологом *Heijiro Tsuji* [6]. На русский язык этот опросник переведён А.Б. Хромовым [7].

Некоторые исследователи считают, что модель личностных черт «Большая пятёрка» можно интегрировать в современные психиатрические модели [5, 8-10].

Анализ больших данных и поведения пользователей социальных сетей открывает новые возможности для исследования личностных черт, такие как построение и проверка предсказательных моделей о личностных чертах и поведении людей в норме и патологии, в том числе и с использованием русскоязычных данных [11]. По качеству сбор данных различных форматов сопоставим с данными, собранными в режиме реального времени. Такая процедура сбора данных позволяет значительно увеличить размеры выборки [12].

В данной работе предлагается алгоритм психолингвистического анализа данных социальных сетей с применением методов обработки текстов на естественном языке и машинного обучения при использовании модели оценки психологических особенностей человека «Большая пятёрка».

1 Обзор предшествующих работ по исследованию личностных характеристик пользователей социальных сетей

Возможность определения личностных характеристик пользователя на основе слабоструктурированной информации рассматривалась в различных работах.

В статье [13] исследование личностей авторов различных блогов проводилось с помощью психологического опроса. Было показано, что использование некоторых слов может быть связано с личностными характеристиками автора. В [14] утверждается, что используемые слова и структура текста могут отражать те или иные черты личности автора блога.

В исследовании [15] авторы опросили более 70 пользователей, имеющих страницы в социальных сетях, и провели анализ текстовых заметок этих пользователей. При помощи метода опорных векторов (*support vector machine, SVM*) и выделения *N*-грамм из текстов была показана возможность определения личностных характеристик.

В статье [16] по результатам анализа текстов пользователей социальной сети *Twitter* показано, что можно оценить личность автора по его текстовым заметкам, а также с помощью дополнительной информации, такой как количество слов в сообщении, количество подписчиков страницы и т.д.

В работе [17] приведены результаты анализа пользователей социальной сети. Ста пользователям было предложено пройти опрос на основе пятифакторной модели личности. С помощью соответствующего *API* была получена информация со страниц этих пользователей. На основе методов интеллектуального анализа данных были определены личностные характеристики пользователей.

В работах [18-21] показана возможность применения методов интеллектуального анализа данных при работе с изображениями с целью получения личностных характеристик пользователя.

Большинство работ в этой области проводится на основе англоязычной текстовой информации, а в работах [22, 23] выполнен анализ текстов из социальных сетей русскоязычных сообществ.

2 Алгоритм психолингвистического анализа данных социальных сетей

Психолингвистический анализ относится к задаче бинарной классификации по пяти факторам. Основой предлагаемого подхода является решение задачи классификации текстов пользователей социальных сетей с целью определения психолингвистических характеристик автора. Общая схема классификации включает несколько последовательных этапов (рисунок 1).

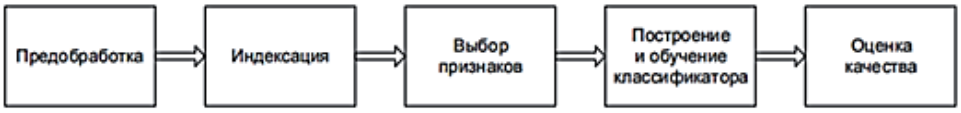


Рисунок 1 – Общая схема классификации

Этапы построения и обучения классификатора, обеспечивающего психолингвистический анализ данных социальных сетей с использованием машинного обучения и модели «Большая пятёрка», представлены на рисунке 2.

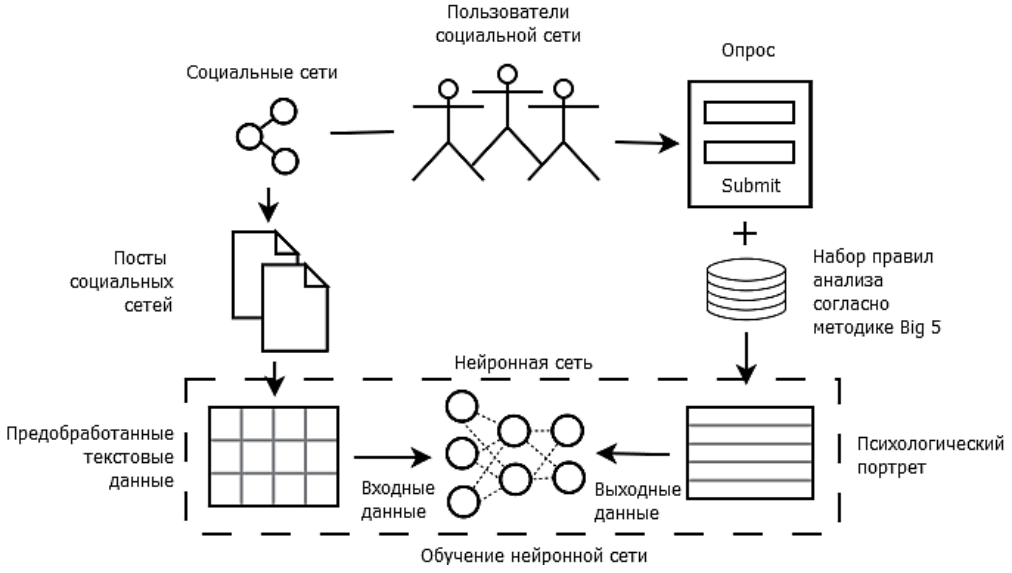


Рисунок 2 – Этапы построения и обучения классификатора

В качестве входных данных используются тексты из «постов» страниц пользователей в социальных сетях. Используются только тексты, написанные пользователем социальной сети, копии текстов других страниц («репосты») не учитываются. Тексты содержат личные мнения, рассуждения и мысли авторов. Основные источники текстов на страницах:

- сообщения («посты»);
- текстовые статусы;
- комментарии к собственным и иным сообщениям.

Размеры текстов варьируются от одного-двух предложений до нескольких десятков предложений.

Для сбора выходных данных обучающей выборки пользователям социальных сетей, тексты которых анализируются, предлагается пройти тест *5PFQ* [7]. В данном тесте 75 вопросов, и он определяет выраженность 30 черт личности (5 основных факторов и 25 первичных). Опросник содержит парные высказывания, опрашиваемый определяет степень собственного согласия с тем или иным высказыванием по шкале от 0 до 4.

Пример вопроса: «Всё новое вызывает у меня интерес» - «Часто новое вызывает у меня раздражение». Все вопросы определены разработчиками опросника согласно модели «Большая пятёрка» [6, 7].

По результатам проведённого теста для каждого пользователя определяются значения психологических характеристик для пяти факторов.

Пример выходных значений для фактора эмоциональности приведён в таблице 1. Выходные значения для всех пяти факторов эмоциональности также определены разработчиками опросника согласно модели «Большая пятёрка» [6, 7].

Таблица 1 – Выходные значения фактора эмоциональности.

Повышенная эмоциональность:	Эмоциональная сдержанность:
Люди чувствуют себя беспомощными, неспособным справиться с жизненными трудностями. Их поведение во многом обусловлено ситуацией. Они с тревогой ожидают неприятностей, в случае неудачи легко впадают в отчаяние и депрессию. Такие люди хуже работают в стрессовых ситуациях, в которых испытывают психологическое напряжение. У них, как правило, занижена самооценка, они обидчивы и в неудачах, в первую очередь, обвиняют себя.	Низкие значения по этому фактору свойственны лицам самодостаточным, уверенным в своих силах, эмоционально зрелым, смело смотрящим в лицо фактам, спокойным, постоянным в своих планах и привязанностях, не поддающимся случайным колебаниям настроения. На жизнь такие люди смотрят серьезно и реалистично, хорошо осознают требования действительности, не скрывают от себя собственных недостатков, не расстраиваются из-за пустяков, чувствуют себя хорошо приспособленными к жизни.

Полученные значения психологических характеристик пользователей составляют выходные значения для нейронной сети и соответствуют следующей модели:

$$Out = \{N, E, O, A, C\}, \quad (1)$$

где: N – Нейротизм;
E – Экстраверсия;
O – Открытость опыту;
A – Согласие, или сотрудничество;
C – Сознательность, или добросовестность.

Так как, согласно модели «Большая пятёрка», по каждому фактору эмоциональности опрашиваемые разбиваются на два класса, то $|N|=|E|=|O|=|A|=|C|=2$.

Для определения психологических особенностей пользователя на вход обученному классификатору подаются текстовые данные со страницы пользователя социальной сети, а на выходе получаются пять психологических характеристик по одной для каждого из пяти факторов.

Тексты социальных сетей имеют ряд особенностей, затрудняющих процесс их семантического анализа. Это наличие грамматических ошибок, эмотиконов (иконка с эмоцией, от англ. *emoticon, emotionicon*) и др.

Предобработка постов пользователей социальной сети включает следующие этапы.

1) графематический анализ текста (разбиение текста на простые предложения; обычно используется не больше трёх предложений);
2) исправление орфографических ошибок (в данной работе использовалась онтология *RuWordNet* [24]);
3) удаление стоп-слов.
4) лемматизация слов и словосочетаний (в данной работе использовалась система *mystem* [25]).

В качестве метода индексации использовался метод *N*-грамм, для выбора признаков – метод *TF-IDF* [26].

В качестве классификаторов использовались два метода.

Метод опорных векторов - это метод линейной классификации. Модель SVM подробнее описана в работе [27].

Главными преимуществами метода опорных векторов являются:

- высокое качество;
- возможность работы с небольшим набором данных для обучения;
- сводимость к задаче выпуклой оптимизации, имеющей единственное решение.

Метод случайного леса (*RF*, от англ. *Random Forest*) относится к методам логической классификации. В работе использовалась модель классификатора, описанная в [28]. Основные преимущества метода: простая программная реализация, понятность и интерпретируемость получаемых результатов.

3 Результаты экспериментов

В задачах бинарной классификации для оценки качества часто применяют метрику площади под кривой ошибок (*AUC ROC, Area Under ROC Curve, Receiver Operating Characteristic*) [29]. Кривая ошибок показывает зависимости *FPR* (от англ. *False Positive Rate*) и *TPR* (от англ. *True Positive Rate*). *FPR*– это процент точек (объектов) 1-ого класса, которые неверно классифицированы алгоритмом, *TPR* – это процент точек (объектов) 2-ого класса, которые верно классифицированы алгоритмом. В этих координатах (*FPR, TPR*) строится *ROC*-кривая. Площадь под кривой ошибок является характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше площадь, тем лучше классификация.

В рамках проведения экспериментов оценивалась эффективность алгоритмов классификации текстов с целью определения психолингвистических характеристик пользователя социальной сети. Оцениваемые алгоритмы были основаны на двух моделях классификаторов (*SVM* и *RF*), а также включали разделение классифицируемых объектов на обучающую и тестовую выборки в различных соотношениях.

Эксперименты проводились на выборке из участников психологического опроса по модели «Большая пятёрка», содержащей загруженные данные профилей участников. Общая схема проведения экспериментов представлена на рисунке 3.

Разработана программная система, которая включает следующие компоненты (и их назначение):

- онтология *RuWordNet* (для исправления грамматических ошибок в анализируемых текстах);

- библиотека *nltk.corpus.stopwords* (для удаления стоп-слов);
- система *mystem* (для лемматизации текстов);
- библиотека *scikit-learn.CountVectorizer* (для индексации текстов с использованием алгоритма *N*-грамм);
- библиотека *scikit-learn.TfidfTransformer* (для определения *TF-IDF* меры);
- библиотека *scikit-learn.GridSearchCV* (для выбора параметров алгоритма классификации);
- библиотека *sklearn.metrics* (для получения метрик точности, полноты и правильности).

Для экспериментов использовались открытые текстовые данные со страниц 1126 пользователей социальной сети.

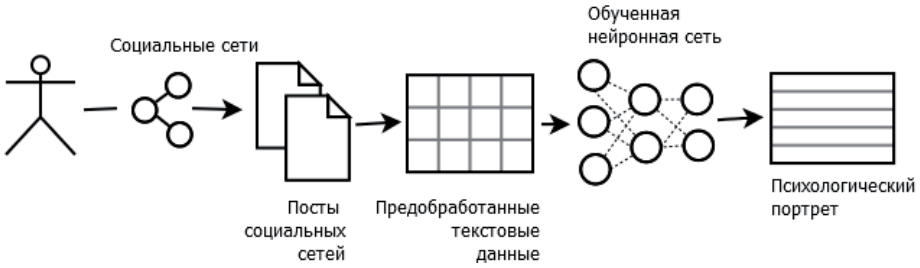


Рисунок 3 – Схема проведения экспериментов

Результаты классификации оценивались по метрике *AUC ROC*. Для оценки было произведено пять итераций алгоритма со случайным сбалансированным распределением элементов по классам в обучающей и тестовой выборках. На каждой итерации оптимальные параметры алгоритма классификации подбирались методом перебора.

Было проведено три множества экспериментов с разбивкой множества классифицируемых объектов на обучающую и тестовую выборки в соотношениях: 70/30, 60/40 и 50/50. Результаты экспериментов представлены на рисунке 4.

Как видно из рисунка, наименьшую эффективность предложенный подход показал при определении объектов классов «*Согласие*» и «*Открытость опыту*». Это связано с большим дисбалансом по классам для данных характеристик в исходной выборке. Методом SVM получены несколько лучшие результаты для классов «*Сознательность*» и «*Нейротизм*». В исследованиях [8-12] были получены схожие результаты.

Заключение

Предложен подход к определению психологических характеристик пользователя социальных сетей посредством анализа текстовых сообщений в социальных сетях с использованием машинного обучения. Новизна подхода заключается в использовании в качестве обучающих и тестовых данных для классификатора результатов прохождения пятифакторного опросника личности, полученных на основе предобработанных текстовых данных со страниц социальных сетей пользователя.

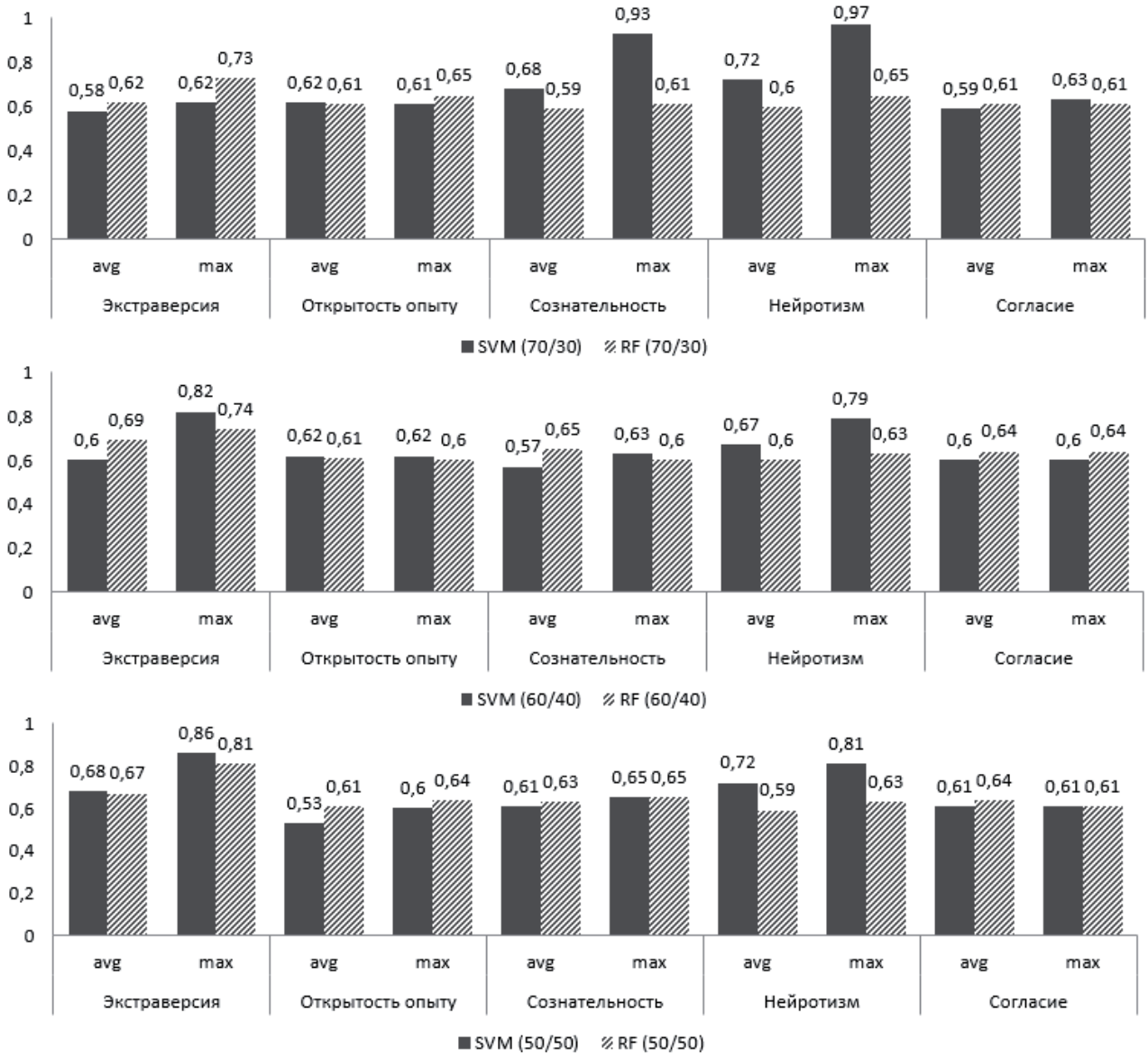


Рисунок 4 – Результаты экспериментов

Список источников

[1] ***Польская Н.А., Якубовская Д.К.*** Влияние социальных сетей на самоповреждающее поведение у подростков. *Консультативная психология и психотерапия*. 2019. Т. 27. № 3. С.156-174. DOI:10.17759/cpp.20192703010.

[2] ***Наумов В.В.*** Анализ социальной структуры интернет-аудитории. *Вестник Челябинского государственного университета*. 2012, Т.35(289). С.148-153.

[3] ***Хайтун С.Д.*** Количественный анализ социальных явлений. Изд. 3-е, КомКнига. М.: 2010. 280 с.

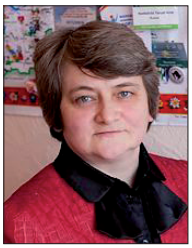
[4] ***Widiger T.A.,*** Mullins-Sweatt S.N. Clinical utility of a dimensional model of personality disorder. *Professional Psychology: Research and Practice*, 2010; 41(6): 488-494.

[5] ***Widiger T.A., Costa P.T., McCrae R.R.*** A proposal for Axis II: Diagnosing personality disorders using the five-factor model. In P.T. Costa, Jr. & T.A. Widiger (Eds.), Personality disorders and the five-factor model of personality. 2002. P.431-456. Washington, DC, US: American Psychological Association. DOI:10.1037/10423-025.

[6] ***Fujishima Y., Yamada N., Tsuji H.*** Construction of Short form of Five Factor Personality Questionnaire. *The Japanese Journal of* Personality, 2004, Volume 13, Issue 2, P.231-241.

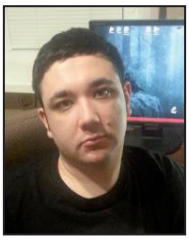
[7] ***Хромов А.Б.*** Пятифакторный опросник личности. Курган: Изд-во Курганского гос. ун-та. 2000. 23 с.
[8] ***Wiggins J.S., Pincus A.L.*** Conceptions of personality disorders and dimensions of personality. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1989; 1(4), 305-316. DOI:10.1037/1040-3590.1.4.305.
[9] ***Piedmont R.L., Sherman M.F., Sherman N.C., Dy-Liacco G.S., Williams J.E.*** Using the five-factor model to identify a new personality disorder domain: the case for experiential permeability. Journal of Personal Social Psychology, 2009. Vol. 96, P.1245-1258.
[10] ***Ozer D.J., Benet-Martinez V.*** Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 2006; Vol. 57, P.401-421.
[11] ***Ледовая Я.А., Боголюбова О.Н., Тихонов Р.В.*** Стресс, благополучие и темная триада. *Психологические исследования*. 2015. Т. 8. № 43. С. 5.
[12] ***Ледовая Я.А., Тихонов Р.В., Боголюбова О.Н.*** Социальные сети как новая среда для междисциплинарных исследований поведения человека. *Вестник Санкт-Петербургского университета. Психология и педагогика*. 2017. Т. 7. № 3. С. 193-210.
[13] ***Yarkoni T.*** Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*. 2010. vol. 44. N 3. P.363-373.
[14] ***Iacobelli F.*** Large scale personality classification of bloggers. International conference on affective computing and intelligent interaction. Springer, Berlin, Heidelberg, 2011. P.568-577.
[15] ***Oberlander J., Nowson S.*** Whose thumb is it anyway? Classifying author personality from weblog text. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. 2006. P.627-634.
[16] ***Golbeck J.*** Predicting personality from twitter. 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE, 2011. P.149-156.
[17] ***Souri A., Hosseinpour S., Rahmani A.M.*** Personality classification based on profiles of social networks' users and the five-factor model of personality. Human-centric Computing and Information Sciences. 2018. v.8. N.1. P.24.
[18] ***Cristani M.*** Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. Proceedings of the 21st ACM international conference on Multimedia. 2013. P.213-222.
[19] ***Segalin C.*** The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. IEEE Transactions on Affective Computing. 2016. vol.8. N.2. P.268-285.
[20] ***Segalin C., Cheng D.S., Cristani M.*** Social profiling through image understanding: Personality inference using convolutional neural networks. Computer Vision and Image Understanding. 2017. vol.156. P.34-50.
[21] ***Steele Jr F.*** Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement. Third International AAAI Conference on Weblogs and Social Media. 2009.
[22] ***Yarushkina N., Filippov A., Moshkin V., Namestnikov A., Guskov G.*** The social portrait building of a social network user based on semi-structured data analysis. CEUR Workshop Proceedings/ 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction, 2019. Vol.2475, 2019, P.119-129.
[23] ***Filippov A., Moshkin V., Guskov G., Romanov A.*** Intelligent Instrumentation for Opinion Mining in Social Media. Proceedings of the II International Scientific and Practical Conference "Fuzzy Technologies in the Industry – FTI 2018". Ulyanovsk, Russia, 23-25 October, 2018. P.50-55.
[24] ***Loukachevitch N., Lashevich G.*** Multiword expressions in Russian Thesauri RuThes and RuWordNet. Proceedings of the AINL FRUCT 2016, 2016. P.66-71.
[25] ***Droganova K.*** Building a dependency parsing model for Russian with maltparser and Mystem tagset. International Workshop on Treebanks and Linguistic Theories (TLT14). 2015. P.268.
[26] ***Ramos J. et al.*** Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. 2003. vol.242. P.133-142.
[27] ***Cauwenberghs G., Poggio T.*** Incremental and decremental support vector machine learning. Advances in neural information processing systems. 2001. pp.409-415.
[28] ***Breiman L.*** Random Forests. Machine Learning: journal. 2001. Vol.45, no.1. P.5-32.
[29] ***Narkhede S.*** Understanding AUC-ROC Curve. Towards Data Science. 2018. vol. 26.

Сведения об авторах



Ярушкина Надежда Глебовна (1962 г. рождения). Доктор технических наук, профессор, ректор Ульяновского государственного технического университета (УлГТУ). Член Российской и Европейской ассоциаций искусственного интеллекта. Область научных интересов - мягкие вычисления, нечёткая логика, гибридные системы. Опубликовано более 390 научных работ. Author ID (RSCI): 10358; Author ID (Scopus): 6602353202; Researcher ID (WoS): B-4438-2014; ORCID: 0000-0002-5718-8732.*jng@ulstu.ru*.

Мошкин Вадим Сергеевич (1990 г. рождения). Окончил УлГТУ в 2012 г., к.т.н. (2017 г.), доцент кафедры «Информационные системы» УлГТУ. Директор Департамента цифровой трансформации УлГТУ. Член Российской и Европейской ассоциаций искусственного интеллекта. В списке научных трудов более 150 статей в области интеллектуальной обработки знаний, автоматизации проектирования, а также построения прикладных интеллектуальных систем. Author ID (RSCI): 762084; Author ID (Scopus): 57190250573; Researcher ID (WoS): L-3578-2016; ORCID: 0000-0002-9258-4909. *v.moshkin@ulstu.ru*. ✉.



Андреев Илья Алексеевич (1994 г. рождения). Окончил УлГТУ в 2017 г., аспирант кафедры «Информационные системы» УлГТУ. Заведующий лабораторией автоматизации образовательного процесса УлГТУ. Имеет более 40 статей в области автоматизации проектирования, онтологического инжиниринга и технической лингвистики.Author ID (RSCI): 842148; Author ID (Scopus): 57190248754; ORCID: 0000-0002-6217-9566.*ares-ilya@yandex.ru*.

Поступила в редакцию 14.01.2022, после рецензирования 28.03.2022. Принята к публикации 30.03.2022.

Algorithm for psycholinguistic analysis of social networks texts using the Big Five Personality Traits

© 2022, N.G. Yarushkina, V.S. Moshkin ✉, I.A. Andreev

Ulyanovsk State Technical University, Ulyanovsk, Russia

Abstract

The paper presents an approach to determining the psychological characteristics of a user of social networks through the analysis of text messages in social networks. The proposed approach includes the user's texts classification using machine learning. The results of the analysis of user surveys in accordance with the Big Five model, as well as a set of author's text data from social network pages, are used as training data. The questionnaire contains paired statements, and the respondent determines the degree of their own agreement with one or another statement on a scale from 0 to 4. Natural language text processing (NLP) methods were applied to the text resources used as input data for the classifier, as well as the RuWordNet linguistic ontology, in order to level out a number of features of social network texts, for example, the presence of grammatical errors and emoticons that complicate the process. semantic analysis. Two models were used as classifiers: the support vector machine and the random forest method. The area under the error curve (AUC ROC) metric was used to evaluate performance. The experiments used open text data of more than 1000 users of social networks.

Key words: *Big Five Model, machine learning, social network, Psycholinguistic Analysis.*

Citation: *Yarushkina NG, Moshkin VS, Andreev IA.* Algorithm for psycholinguistic analysis of social networks texts using the Big Five Personality Traits [In Russian]. *Ontology of designing*. 2022; 12(1): 82-92. DOI:10.18287/2223-9537-2022-12-1-82-92.

Financial Support: This work was supported Ministry of Education and Science of Russia in framework of project № 075-00233-20-05 from 03.11.2020 «Research of intelligent predictive multimodal analysis of big data, and the extraction of knowledge from different sources».

Conflict of interest: The author declares no conflict of interest.

List of figures and tables

References

[1] ***Polskaya NA, Yakubovskaya DK.*** The influence of social networks on self-harming behaviour on adolescents. Consultative psychology and psychotherapy. 2019; 27(3): 156-174. DOI:10.17759/cpp.20192703010.
[2] ***Naumov VV***. Analysis of the social structure of the Internet audience [In Russian]. Bulletin of the Chelyabinsk State University, 2012; 35(289): 148-153.
[3] ***Khaitun, S.D.*** Quantitative analysis of social phenomena [In Russian]. Ed. 3rd, ComBook. Moscow: 2010. 280 p.
[4] ***Widiger, T.A.,*** Mullins-Sweatt S.N. Clinical utility of a dimensional model of personality disorder / Professional Psychology: Research and Practice, 2010; 41(6): 488-494.
[5] ***Widiger TA, Costa PT, McCrae RR.*** A proposal for Axis II: Diagnosing personality disorders using the five-factor model. In P.T. Costa, Jr. & T.A. Widiger (Eds.), Personality disorders and the five-factor model of personality (pp.431-456). Washington, DC, US: 2002, American Psychological Association. DOI:10.1037/10423-025.
[6] ***Fujishima Y, Yamada N, Tsuji H.*** Construction of Short form of Five Factor Personality Questionnaire, The Japanese Journal of Personality, 2004; 13(2): 231-241.
[7] ***Khromov AB.*** Five-factor personality questionnaire [In Russian]. Kurgan: Publishing House of the Kurgan State. university. 2000. 23 p.
[8] ***Wiggins JS, Pincus AL.*** Conceptions of personality disorders and dimensions of personality. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1989; 1(4): 305-316. DOI:10.1037/1040-3590.1.4.305.
[9] ***Piedmont RL, Sherman MF, Sherman NC, Dy-Liacco GS, Williams JE.*** Using the five-factor model to identify a new personality disorder domain: the case for experiential permeability. Journal of Personal Social Psychology, 2009; 96: 1245-1258.
[10] ***Ozer DJ, Benet-Martinez V.*** Personality and the prediction of consequential outcomes. Annual Review of Psychology, 2006; 57: 401-421.
[11] ***Ledovaya YaA., Bogolyubova ON., Tikhonov RV.*** Stress, well-being and the dark triad. *Psychological research*. 2015; 8(43): 5.
[12] ***Ledovaya YA, Tikhonov RV, Bogolyubova ON.*** Social networks as a new environment for interdisciplinary studies of human behavior. Vestnik of Saint Petersburg University. Psychology, 2017; 7(3): 193–210. DOI:10.21638/11701/spbu16.2017.301.
[13] ***Yarkoni T.*** Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. Journal of research in personality. 2010; 44(3): 363-373.
[14] ***Iacobelli F.*** Large scale personality classification of bloggers. International conference on affective computing and intelligent interaction. Springer, Berlin, Heidelberg, 2011. P.568-577.
[15] ***Oberlander J, Nowson S.*** Whose thumb is it anyway? Classifying author personality from weblog text. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. 2006. P.627-634.
[16] ***Golbeck J.*** Predicting personality from twitter. 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE, 2011. P.149-156.
[17] ***Souri A, Hosseinpour S, Rahmani AM.*** Personality classification based on profiles of social networks' users and the five-factor model of personality. Human-centric Computing and Information Sciences. 2018; 8(1): 24.
[18] ***Cristani M.*** Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. Proceedings of the 21st ACM international conference on Multimedia. 2013. P.213-222.
[19] ***Segalin C.*** The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. IEEE Transactions on Affective Computing. 2016; 8(2): 268-285.

[20] ***Segalin C, Cheng DS, Cristani M.*** Social profiling through image understanding: Personality inference using convolutional neural networks. Computer Vision and Image Understanding. 2017; 156: 34-50.
[21] ***Steele JrF et al.*** Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement. Third International AAAI Conference on Weblogs and Social Media. 2009.
[22] ***Yarushkina N, Filippov A, Moshkin V, Namestnikov A, Guskov G.*** The social portrait building of a social network user based on semi-structured data analysis. CEUR Workshop Proceedings. 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction, IS 2019. 2019; 2475: 119-129.
[23] ***Filippov A, Moshkin V, Guskov G, Romanov A.*** Intelligent Instrumentation for Opinion Mining in Social Media. Proceedings of the II International Scientific and Practical Conference "Fuzzy Technologies in the Industry – FTI 2018". Ulyanovsk, Russia, 23-25 October, 2018. P.50-55.
[24] ***Loukachevitch N, Lashevich G.*** Multiword expressions in Russian Thesauri RuThes and RuWordNet. Proceedings of the AINL FRUCT 2016, 2016. P.66-71.
[25] ***Droganova K.*** Building a dependency parsing model for Russian with maltparser and Mystem tagset. International Workshop on Treebanks and Linguistic Theories (TLT14). 2015. P.268.
[26] ***Ramos J. et al.*** Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. 2003; 242: 133-142.
[27] ***Cauwenberghs G, Poggio T.*** Incremental and decremental support vector machine learning. Advances in neural information processing systems. 2001. P.409-415.
[28] ***Breiman L.*** Random Forests. Machine Learning: journal. 2001; 45(1): 5-32.
[29] ***Narkhede S.*** Understanding AUC-ROC Curve. Towards Data Science. 2018. vol. 26.

About the authors

Nadezhda Glebovna Yarushkina (b. 1962) Doctor of Technical Sciences, Professor, the Rector of Ulyanovsk State Technical University, a Member of the Russian and European Association of Artificial Intelligence. Her research interests include soft computing, fuzzy logic, and hybrid systems. Published over 390 scientific papers. Author ID (RSCI): 10358; Author ID (Scopus): 6602353202; Researcher ID (WoS): B-4438-2014; ORCID: 0000-0002-5718-8732. *jng@ulstu.ru.*

Vadim Sergeevich Moshkin (b. 1990) graduated from Ulyanovsk State Technical University (UlSTU) in 2012, got the Ph. D. in 2017, associate professor of the Information Systems department at UlSTU. Director of Digital Transformation Department at UlSTU. Member of the Russian and European Association of Artificial Intelligence. He is a co-author of more than 150 publications in the field of data mining, design automation and construction of applied intelligent systems. Author ID (RSCI): 762084; Author ID (Scopus): 57190250573; Researcher ID (WoS): L-3578-2016; ORCID: 0000-0002-9258-4909. *v.moshkin@ulstu.ru*. ✉.

Ilya Alekseevich Andreev (b. 1994) graduated from Ulyanovsk State Technical University (UlSTU) in 2017, a graduate student of the Information Systems department of UlSTU. Head of the laboratory of educational process automation UlSTU. He is a co-author of more than 40 articles in the field of design automation, ontological engineering and technical linguistics. Author ID (RSCI): 842148; Author ID (Scopus): 57190248754; ORCID: 0000-0002-6217-9566. *ares-ilya@yandex.ru*.

Received January 14, 2022. Revised March 28, 2022. Accepted March 30, 2022.