Computing, Telecommunication and Control, 2023, Vol. 16, No. 3, Pp. 54–63.
Информатика, телекоммуникации и управление. 2023. Том 16, № 3. С. 54–63.

Intellectual Systems and Technologies
Интеллектуальные системы и технологии

Research article
DOI: https://doi.org/10.18721/JCSTCS.16305
UDC 004

ANALYSIS OF PERSONALITY TRAITS BASED ON THE DISC MODEL USING MACHINE LEARNING METHODS

L.P. Mbele Ossiyi ✉ , *P.D. Drobintsev*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ lucprucell@gmail.com

Abstract. The analysis of a person's social media behavior with respect to privacy and human rights provides information about their personality traits and is seen as a topical task today. In areas such as marketing, training, education, human resource management and hiring policies in companies, knowledge about personality traits proves to be profitable and important in decision making and business orientation cases. The paper analyzes the performance of machine learning methods in a personality trait identification task based on the DISC psychological model and a small size dataset created from scratch. Although the dataset created was relatively small, the machine learning methods used showed encouraging and convincing results. Results for all personality trait classifiers were improved using hyperparameter optimization, increasing the performance of the XGBoost classifier to 70.45% on the accuracy metric in the test sets.

Keywords: DISC model, personality traits, data pre-processing, machine learning, TF-IDF, XGBoost

Citation: Mbele Ossiyi L.P., Drobintsev P.D. Analysis of personality traits based on the disc model using machine learning methods. Computing, Telecommunications and Control, 2023, Vol. 16, No. 3, Pp. 54–63. DOI: 10.18721/JCSTCS.16305

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.16305
УДК 004

АНАЛИЗ ЛИЧНОСТНЫХ ЧЕРТ НА ОСНОВЕ МОДЕЛИ DISC С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Л.П. Мбеле Оссийи ✉ *, П.Д. Дробинцев*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ lucprucell@gmail.com

Аннотация. Анализ поведения человека в социальных сетях с соблюдением конфиденциальности и прав человека позволяет получить информацию о его личностных чертах и рассматривается на сегодняшний день как актуальная задача. В таких сферах как маркетинг, профессиональная подготовка, образование, управление человеческими ресурсами и политика найма в компаниях, знание о личностных чертах оказывается прибыльным и важным в случаях принятия решений и ориентации на бизнес. Статья посвящена анализу производительности методов машинного обучения в задаче идентификации личностных черт на основе психологической модели DISC и созданного с нуля набора данных небольшого размера. Хотя созданный набор данных был относительно небольшого размера, используемые методы машинного обучения показали обнадеживающие и убедительные результаты. Результаты, полученные всеми классификаторами по всем чертам личности, были улучшены с применением оптимизации гиперпараметров, что позволило увеличить производительность классификатора XGBoost до 70,45% по метрике accuracy в тестовых наборах.

Ключевые слова: модель DISC, личностные черты, предварительная обработка данных, машинное обучение, TF-IDF, XGBoost

Для цитирования: Mbele Ossiyi L.P., Drobintsev P.D. Analysis of personality traits based on the disc model using machine learning methods // Computing, Telecommunications and Control. 2023. Т. 16, № 3. С. 54–63. DOI: 10.18721/JCSTCS.16305

Introduction

The last two decades have witnessed massive use of digital platforms for sharing ideas, feelings, opinions and emotions, and X (former Twitter which is banned in Russia Federation) is one of the most widely used platforms in this context. Automatic identification of users' personality traits based on publicly available information from online social platforms is increasingly becoming a promising field these days and attracts widespread interest in various disciplines [1, 2]. As of today, automatic identification of personality traits is likely to become beneficial in many areas, such as recommendation systems, attribution of authorship, implementation of recruitment policies. Such applications are mainly built based on psychological models such as MBTI, Big Five, DISC and usually require large datasets, machine and deep learning algorithms. The size of a dataset is an important component in determining the performance of a machine-learning model. Large datasets generally lead to better performance on the classification problem. Nevertheless, it is stated that, large datasets in personality traits identification tasks are mostly created in English language and it is quite difficult to find a dataset for a specific language, which may lead to use of small dataset. In machine learning, a small dataset usually refers to a dataset containing less than 1000 instances while calculations and mathematical computations on it can be performed on a computer in a short time. Such a dataset is considered small because it may not be representative of the population. Some previous works have been devoted to the identification of personality traits based on the psychological models Big Five [3], DISC [4, 5] and MBTI [6]. However, the existing datasets used in these papers are mostly outdated,

in English and unbalanced, making it difficult to make assumptions about different languages. An unbalanced dataset is one in which samples and their corresponding labels are not evenly distributed over the data space [7]. The unbalanced data distribution problem occurs when majority classes have a larger proportion of features than minority classes.

To address this problem, a balanced dataset of small size, in French, based on the DISC psychological model was created and the results and performance of machine learning algorithms on this dataset were analyzed.

In this paper, a dataset in French language was used, which is less difficult and different for instance from Russian language from a linguistic point of view (Cyrillic and Latin alphabet). Russian language, like other Slavic languages is a morphologically rich language with free order and inflection. These linguistic factors make it difficult to collect enough relevant features data to efficiently train machine-learning models, which could produce lower quality results compared to French ones.

The choice of the DISC model is explained by its ability to be a predictor of improved manageability in work teams (enterprises, organizations) [8]. Thus, the present paper aims to analyze machine-learning methods performance on a balanced small dataset in the task of personality traits identification based on the DISC psychological model.

Problem statement

To achieve our objective, this paper is organized as follows:

1. the dataset creation in French and data annotation;
2. data pre-processing;
3. application of different machine learning algorithms and analysis of their performance.

The social network Twitter (Twitter API) [9] was used to collect user-related data in this work. The data (tweets) were collected from January to May 2022 in French.

The psychological model used within the work is DISC. Personality traits were divided into four classes: *dominance, influence, steadiness, consciousness*.

As a result, data were collected on 660 users and more than 144,117 tweets. The corpus was divided into training and test samples in the 80/20 ratio.

The next steps after data collection were cleaning the dataset, its preprocessing and annotation, which was made possible largely by contacting an online service specializing in data annotation.

The following technology stack was used to implement the algorithms: Python 3.9 programming language; NumPy, Matplotlib, SciPy libraries; Google Colab development environment.

Stages of work

In this paper, our main concern was focused on user privacy in the process of collecting and analyzing data from Twitter accounts. The main challenge was to respect the boundary between public and private information. Anonymity of user data was protected by replacing usernames with some codes. Sensitive data, such as age, gender or civil identity of users were neither published nor used in the selection, classification, and other processes in this work.

a) Dataset creation

During the dataset creation process, the following were used: Twitter API, keywords. To obtain relevant data from users, keywords related to emotion, derived from Watson and Tellegen two-dimensional emotion map (1985) were selected. The Tellegen-Watson model (Fig. 1) is useful for linking spatial and discrete levels of emotions [10].

The dotted lines are the upper-level dimensions. The dimensions of positive effect and negative effect are shown as solid lines from the middle of the hierarchy and provide the heuristics needed to distinguish specific words with discrete emotions based on function. Discrete emotions near the axis correlate strongly with this dimension [10].

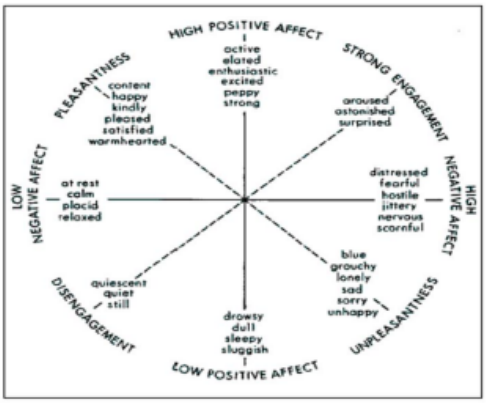


Fig. 1. Two-dimensional map of emotion by Wilson and Tellegen

To analyze the personality traits of individuals and build up a clean dataset consisting only of personal and real accounts, it was essential during the collection process for a better analysis to delete tweets containing languages other than French in their structure and to detect and remove bots and typical accounts such as: fanatical (sport, music), poetry and thoughts ones. During the filtering process several numerical characteristics were considered, such as the number of followed people, the number of friends, the frequency of posted tweets, the frequency of retweets, the frequency of comments and replies, and the age of the account (when the account was created). These characteristics were used by the Support Vector Machine algorithm during the training phase to proceed a binary classification to distinguish real accounts from others. The Support Vector Machine algorithm fitted well for this task due to its ability to handle high-dimensional data. A high value for metrics such as the precision and accuracy of an account indicated a high probability of a non-personal or fake account. For example, sports-related accounts have the particularity of having most of the tweets containing images, videos (taken from other sports-related accounts) with high frequency of replies and retweeting all sporting events with high frequency. As a result, **660** users and **144,117** tweets were obtained as material for work. To ensure work productivity, it was crucial to call on experts in the psychological field to annotate the dataset. The personality traits of the users were assessed by a panel of psychology experts on Fiverr.com. The Fiverr choice as a freelance platform for searching psychological experts is explained by its wide range of services and opportunities for finding services easily. To annotate the dataset, we opted for the choice of several psychological experts to compare the different results and judge their similarity. Psychological experts were chosen based on several criteria such as: their grades, their levels (diplomas and certificates, as Fiverr has a verification process during which sellers provide information about their training and background), and comments from other buyers. To determine user personality traits, each user was labelled with their most frequently used choice of words that fit the model.

Fig. 2 identifies the proportions of personality types: I (influence), D (dominance), C (conscientiousness) and S (steadiness). It is observed that personality types I and D are possessed by most users, and type S by a minority. The annotation work resulted in the following dataset (Fig. 3).

b) Data pre-processing

The obtained data is unstructured (Fig. 4), that is, it contains text, special characters, images, and videos. Hence, some pre-processing steps are required for further processing (Fig. 5) such as: converting all tweets to lowercase to create consistent text; removing stop words such as le, un; removing URLs, emoticons and special characters used in tweets; applying tokenization and lemmatization. Although users can use emoticons in a post to express their feelings and to provide relevant information, however in our case the analysis of users' posts in our dataset showed a low frequency of emoticon used per post, or even per user, consequently in our work we opted to remove them completely for consistency in our analysis.

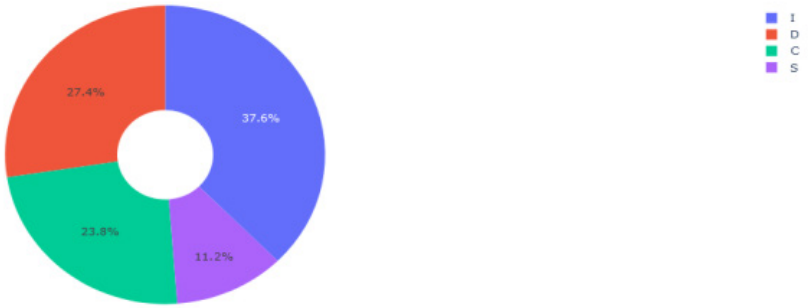


Fig. 2. Percentage of users by traits

Dataframe avec 660 fichiers et 7 colonnes

	namecode	tweets	D	I	S	C	trait
0	cc4de35c353ea3c42fd19afc559669e1	laksa chef heureux centre pailou rue dixon mei...	28.2	28.2	29.4	30.6	C
1	7e3695f6b6fc8252e27e346d78dd89e1	c l espoir soeur sophie mignon 100 x j aimerai...	35.8	27.6	30.8	30.2	D
2	5945481e31aff69d35425af457c1b54b	citation drôle salle d audience mettre face no...	29.6	34.0	25.6	39.0	C
3	6e182efe093be5ba12745c2d11a7592a	laissez savoir préface contact recommande voul...	32.6	24.8	18.2	38.4	C
4	ffd581ce516fe9e2422cb643d63c7e66	boisson gratuite téléchargé l application ipho...	17.6	20.8	23.0	34.4	C
...	...	...	...	...	...	...	...
655	98ba02025b3c0b014b81a63bdfbc61df	faite okayyyy youuu pouvoir ninang haha étrang...	16.6	17.2	23.2	12.6	S
656	459e67ccc3a7873bd23f450b564b4844	effectivement engagez 3 an pouvoir mettre nive...	11.8	10.2	13.8	6.8	S
657	187ad246408231b9b1f1c481f863f4c6	regarder miroir ouai gettin argent l emballage...	8.4	16.2	8.4	19.0	C
658	c9539bdfa985a4f069ee841435a2d424	jose vien hot dogs bière pouvoir détendre fill...	7.2	12.0	12.8	9.0	S
659	4a362c777db5398e0ebf52cc8b25d5c4	prochains jour congé oui j aime l ennemi incro...	9.8	14.2	17.8	8.6	S

660 rows × 7 columns

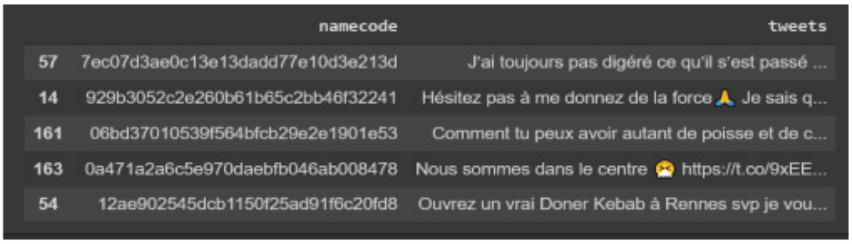
Fig. 3. Annotated dataset

c) Model training

Before starting the model training process, the key point is feature extraction [6], which is the vector representation of texts. In this work, the TF-IDF model [9] was chosen, TF-IDF score is useful to adjust the weight between common and less frequently used words. A data augmentation method by synthetic minority oversampling, SMOTE [2], was applied to balance the dataset. In our work we have trained our dataset with supervised machine learning algorithms corresponding to the classification problem. To have the best algorithm according to several metrics, we used a certain number of classes of algorithms such as logistic regression, decision trees and gradient boosting algorithms.

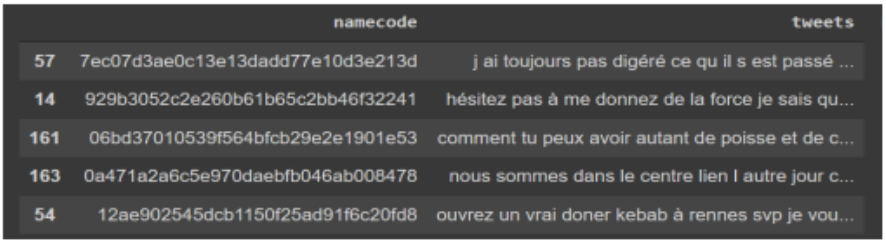
Research results and their discussion

To evaluate the performance of the machine learning models, we used several machine learning metrics such as: accuracy, precision, completeness, error matrix, ROC curve, F1-score. Also, to obtain a model that correctly generalizes the results, we were keen to separate all the training and validation data sets during a time interval T (using temporal separation method). To this end, during the training phase the validation set is unknown to the classifier so that the validation set cannot accidentally infiltrate the training set, and this creates independence between the two sets. To improve the performance of some algorithms, hyperparameter optimization (maximum tree depth, learning rate) was applied [14]. In this paper, the XGBoost algorithm performed the best (Table 1) compared to other machine-learning algorithms. The most important success factors of XGBoost are its scalability in all scenarios and the ability to solve a real-scale problem using a minimum number of resources [15].



	namecode	tweets
57	7ec07d3ae0c13e13dadd77e10d3e213d	J'ai toujours pas digéré ce qu'il s'est passé ...
14	929b3052c2e260b61b65c2bb46f32241	Hésitez pas à me donnez de la force 🙏 Je sais q...
161	06bd37010539f564bfcb29e2e1901e53	Comment tu peux avoir autant de poisse et de c...
163	0a471a2a6c5e970daebfb046ab008478	Nous sommes dans le centre 😭 https://t.co/9xEE...
54	12ae902545dcb1150f25ad91f6c20fd8	Ouvrez un vrai Doner Kebab à Rennes svp je vou...

Fig. 4. Dataset part state before preprocessing



	namecode	tweets
57	7ec07d3ae0c13e13dadd77e10d3e213d	j ai toujours pas digéré ce qu il s est passé ...
14	929b3052c2e260b61b65c2bb46f32241	hésitez pas à me donnez de la force je sais qu...
161	06bd37010539f564bfcb29e2e1901e53	comment tu peux avoir autant de poisse et de c...
163	0a471a2a6c5e970daebfb046ab008478	nous sommes dans le centre lien l autre jour c...
54	12ae902545dcb1150f25ad91f6c20fd8	ouvrez un vrai doner kebab à rennes svp je vou...

Fig. 5. Dataset part state after preprocessing

Table 1

Results of XGBoost application

	Precision	Recall	f1-score	Support
C	0.71	0.55	0.62	31
D	0.69	0.61	0.65	36
I	0.73	0.86	0.79	50
S	0.65	0.73	0.69	15

In Fig. 7, it is observed that each point on the ROC curve is obtained from the values in the error matrix (Fig. 6) associated with applying a certain constraint to the classifier predictions. The ROC curve represents sensitivity as a function of specificity for all possible threshold values of the classifier under study. Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. An excellent model has AUC (Area under curve) near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. The higher the area under the curve, the better the model can perform. For example, our Class C, which represents the lowest population in our dataset has a curve close to the top left and achieve a sensitivity of 80% only if the percentage of misclassified positive examples is about 40% (Fig. 7), which is a good result for a classifier that should have practical applications.

Fig. 8 indicates that the accuracy and completeness curve of Class D shows an acceptable result. An accuracy of about 80% is required to achieve60% completeness. The accuracy and completeness score (otherwise F1-score) for Class I is 0.69. The Class S curve shows a sensitivity to predicted positives values of 50% with a true positive rate of about 60%, which is average. Class C for 50% of predicted positive values shows a rate of almost 70% true positive values. When considering all the results of models based on the accuracy metric (Fig. 9), it is stated that, the model based on naive Bayesian classifier showed the lowest accuracy (**48.48%**), and the model based on XGBoost algorithm showed the best result (**70.45%**).

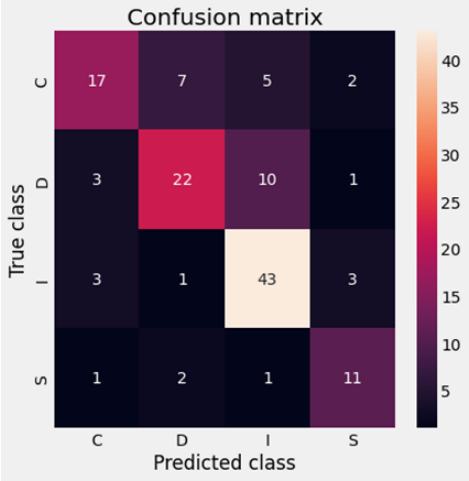


Fig. 6. XGBoost Error matrix

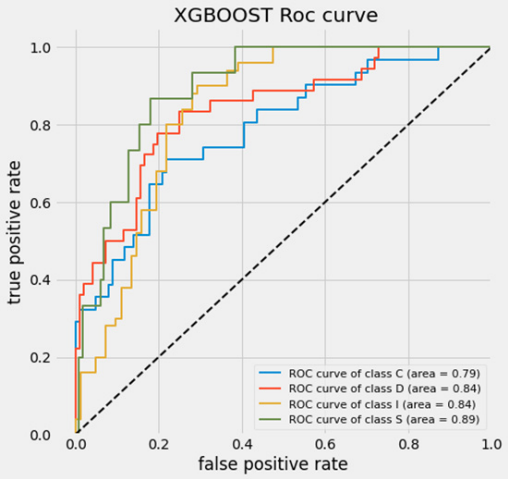


Fig. 7. XGBoost ROC curve

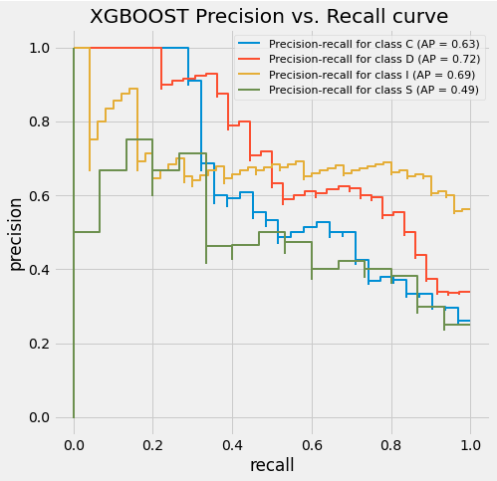


Fig. 8. XGBoost precision and completeness

	Models	Test accuracy
0	XGBoost	70.454545
1	Catboost	68.939394
2	Random Forest	63.636364
3	Logistic regression	63.636364
4	SVM	62.878788
5	SGD Classifier	62.121212
6	Decision Tree classifier	50.757576
7	Multinomial Naive Bayes	48.484848

Fig. 9. Algorithms accuracy

Conclusion

To sum it up, our work showed the feasibility of applying gradient boosting algorithms to the analyzing personality traits task using balanced small dataset. To improve the algorithms performance, hyperparameter tuning was applied to each classifier. Although the accuracy metric is not the only metric for evaluating the performance of a model, the developed model based on the XGBoost algorithm showed the best result according to the accuracy metric (**70.45%**). There is no universal threshold that we use to determine if a model has good accuracy or not. The judgement of good or bad accuracy is subjective and depends on the task in which it is measured. In our case, we obtained a score of 70.45% for a small dataset, so we can consider our model to be useful according to the accuracy metric. In addition, model accuracy between 70% and 90% is realistic and consistent with industry standards.

The following steps were taken to solve the problem:

• Dataset creation in French using the social network Twitter. The dataset was based on the DISC psychological model, which classifies personality traits into 4 categories: dominance, influence, steadiness, conscientiousness.

• Dataset annotation. The dataset was annotated by a team of psychologists on the Fiverr platform. As a result, 660 users and more than 144,000 tweets were selected. After annotating the dataset by a team of psychologists, the first resulting version of the dataset was unbalanced. Thus, the SMOTE data augmentation method was applied for the purpose of balancing the dataset.

• Data preprocessing and model training with several machine learning algorithms. In our paper, some applied pre-processing (semantic, syntactic) and features extraction methods are universal despite the chosen language. Most of the used methods in our work are designed programmatically (based on programming libraries) and support most of the spoken languages of the world.

For further research, there is a need to address the issue of using Russian language material despite the linguistic difficulties of collecting and processing. One of the possible solutions would be to add more data and use other features or extraction methods. The potential of analyzing personality traits using the XGBoost gradient boosting algorithm lies in the possibility of using it in the development of recommendation systems useful in companies, higher education institutions, for marketing, audience targeting, implementation of recruitment policies.

REFERENCES

1. **Xue D., Guo S., Gao L., Wu L.** Personality recognition on social media with label distribution learning. IEEE Access, 2017, Vol.5, pp. 13478–13488.

2. **Setiawan H., Wafi A.A.** Classification of Personality Type Based on Twitter Data Using Machine Learning Techniques. 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020, pp. 94–98.

3. **Goldberg 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. DOI: https://doi.org/10.1037/0022-3514.59.6.1216

4. **Hernández Y., Martínez A., Estrada H., Ortiz J., Acevedo C.** Machine Learning Approach for Personality Recognition in Spanish Texts. Applied Sciences. 2022, Vol. 12, no. 6, pp. 1–17. DOI: https://doi.org/10.3390/app12062985

5. **Dwi Hartanto A., Ema Utami, Sumarni Adi, Suwanto Raharjo.** Classifying User Personality Based on Media Social Posts Using Support Vector Machine Algorithm Based on DISC Approach. 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), Manado, Indonesia, 2020, pp. 1–4.

6. **Bharadwaj S., Sridhar S., Choudhary R., Srinath R.** "Persona Traits Identification based on Myers-Briggs Type Indicator (MBTI) – A Text Classification Approach". International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India. 2018, pp. 1076–1082. DOI: 10.1109/ICACCI.2018.8554828

7. **Tarekegn A.N., Giacobini M., Michalak K.** A review of methods for imbalanced multi-label classification. Pattern Recognition. 2021. Vol. 118. DOI: https://doi.org/10.1016/j.patcog.2021.107965

8. **Chigova E.A., Plyushch I.V., Leskova I.V.** Organization of structured interaction on the base of psychographic characteristics within the model of personality traits DISC. IOP Conference Series Materials Science and Engineering, 2019, Vol. 483, no. 1, pp. 1–6.

9. **Karami A., Morgan Lundy, Frank Webb.** Twitter and Research: A Systematic Literature Review through Text Mining. IEEE Access, 2020, Vol. 8, pp. 67698–67717.

10. **Yang D., Lee W.S.** Disambiguating Music Emotion Using Software Agents. Proceedings of 5th International Conference on Music Information Retrieval, 2004, pp. 1–6.

11. **Goldberg Y.** Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies, 2017, Vol. 37, no. 1, 287 p.

12. **Qaiser S., Ali R.** Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications, 2018, Vol. 181, no. 1, pp. 1–5. DOI : https://doi.org/10.5120/ijca2018917395

13. **Chawla Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, Philip W.** Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, Vol. 16. Pp. 321–357. DOI: https://doi.org/10.1613/jair.953

14. **Agrawal T.** Hyperparameter Optimization in Machine Learning. Apress Berkeley, 2021, 166 p. DOI: https://doi.org/10.1007/978-1-4842-6579-6

15. **Chen T., Guestrin C.** XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. DOI: https://doi.org/10.1145/2939672.2939785

16. **Kluemper D.H., Rosen P.A.** Future employment selection methods: Evaluating social networking web sites. Journal of Managerial Psychology, 2009, Vol. 24, no. 6, pp. 567–580.

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Mbele Ossiyi Luc Prucell
Мбеле Оссийи Люк Прюсель
E-mail: lucprucell@gmail.com

Drobintsev Pavel D.
Дробинцев Павел Дмитриевич
E-mail: drobintsev_pd@spbstu.ru

Submitted: 04.07.2023; Approved: 21.09.2023; Accepted: 11.10.2023.
Поступила: 04.07.2023; Одобрена: 21.09.2023; Принята: 11.10.2023.