The Present Situation and the Prospect of Determining the Personality Type of Text Author with Machine Learning

Ninoslav Čerkez
College for Information Technologies
Zagreb, Croatia
ninoslav.cerkez@vsite.hr

Boris Vrdoljak
University of Zagreb, Faculty of Electrical Engineering and Computing
Zagreb, Croatia
boris.vrdoljak@fer.hr

Sandro Skansi
University of Zagreb, Department for Croatian Studies
Zagreb, Croatia
sskansi@hrstud.hr

***Abstract*— Personality type has an essential role in selecting the best candidates for a job since it gives information regarding the way of thinking and behavior affinities for a candidate. This paper gives an overview of approaching methods before data science, the present situation, and some new directions in this area with an emphasis on determining the personality type of text author in human resource management with machine learning. Since that solving the problem of classification personality type includes using models like MBTI and Big Five, which classify type according to compound types, the focus is on classification problems with compound class labels.**

Keywords—determining personality type, classification, MBTI, Big Five, compound class labels, natural language processing, human resource management, evaluation, and selection.

I. INTRODUCTION

Human resource management is a strategic function in each company, and it faces challenges nowadays since in modern business environment there is a lack of capable candidates and hiring a new capable candidate is essential as finding a new client. At the same time, the term itself has appeared outside the U.S. just after the 1990s, which shows that it is a relatively new view on a function of human resource management [3]. Recruitment and selection are processes inside human resource management processes which covers activities intending to attract the best new candidates for a job. The business study [4] shows how 80% of employee fluctuation has a cause in poor recruitment decisions. Research [5] proves that choosing the wrong candidate has a cost on average of 33% of a worker's annual base salary to replace the wrong candidate and that the cost of turnover will increase 19% from 2017 by 2020, with no signs of changing trends of increasing costs of choosing the wrong candidate. According to the research, the average turnover cost in the USA is $15.000. Society for Human Resource Management survey [6] finds that an average cost-per-hire was $4.425 for non-executive positions and $14.936 for executive positions with increasing trend every year. The last observed increase was 7.1% per year. The survey also states that the duration of the selection process decreased by 16.7%, to 36 days and that recruitment expenses are 15% of all human resource related expenses. The survey also states that only 22% of companies use automated prescreening to review a job applicant's resume and that only 18% of companies use psychometric testing in the process of acquiring new candidates. However, the paper states that this percentage of users who use psychometric testing raises between 10% and 15% per year. These findings imply how considerable is business demand for the improvement of the process of recruitment and that determining personality traits as part of psychometric assessment have an increasing trend in use. Evaluation and selection are challengeable processes in human resource management because there are many parameters which evaluator should include and to complete that task without mistake and fast. Also, in these processes, it is crucial to determine precisely candidate personality type. Since personality is a composite of psychological, physical, and social qualities that distinguish persons from one another and that has properties which enable individuals to interact and to develop themselves, personality assessment aims to identify and evaluate distinctive features of each person and to discover mechanisms which underlie the functioning of personality. Personality assessment procedures through quantitative and qualitative techniques identify psychological qualities, including styles of thinking, feeling, and behaving [7]. There is a diversity in aims, contents, and methods of personality assessment because of different viewpoints, paradigms about personality, and because it is a contemporary field [8]. The variations in observations in behavior styles, affect, and cognition leads to different models of personality determination. Because of their popularity, papers [1] and [2] give a comparison of Myers Briggs and the Big Five personality determination. The five-factor model (FFM) and the OCEAN model are synonyms for the Big Five model.

This paper does not go into the psychological evaluation of personality traits in psychometric assessment since these processes are part of psychology as a science. However, it is possible to give added value with researches in this field. Added value to business resides in cutting costs with shortening the duration and acquired resources in the process of recruitment and selection of candidates. Added value also lays in automation the process and raising the precision of determining personality traits through computer science methods and tools, which can also give a new perspective from the point of computer science tools and methods to other scientific fields, like psychology, and this is the aim of this paper. Information models can help in dealing with processes of evaluation and selection of candidates since these processes nowadays include a considerable amount of data with characteristics of high variety, velocity, and volume and data science tools and methods, like deep learning, can provide insights from such a large amount of data. Without these tools and methods, it would be impossible to learn from nowadays data. There is an also understanding in psychology

that new technologies give extraordinary opportunities to store and to process a large volume of data, to monitor ongoing thoughts, feelings, behaviors, and their biological correlates [7].
Section II covers computer science approaches in selection and evaluation before data science era. Section III covers data science approaches in selection and evaluation candidates for a job. Section IV gives a prospect of researches and trends in selection and evaluation candidates and concludes the paper.

II. Pre-Data-Science Approaches in Selection and Evaluation

Researches are trying nowadays to implement data science methods and approaches in processes in candidate selection and evaluation. However, there are researches which use methods and approaches before data science era.
In situations where there is a certain level of uncertainty, like in processes of evaluation and selection, it is possible to choose the Analytic Hierarchy Process (APH). Saaty created this mathematical model based on linear algebra in 1971, and the process researches completed it in 1980. Initially, the U.S. Army created this process as a part of research projects in the field of making complex decisions. APH suggests problem decomposition in the hierarchy of subproblems for a more honest evaluation, which includes weighted evaluation and measurement [9]. There is also an approach of an upgrade of APH through type-2-fuzzy [17].
The method proposed in [10] covers the model of multiple fuzzy criteriums in making more correct decisions in choosing the candidate for a job.
Ruskova in [11] presents a Decision Support System, which also uses the fuzzy model with more formal methods. This approach does not give significant improvement of APH but mostly gives a demonstration of solving information system.
In paper [12] one demonstrates a process of assigning potential employees through fuzzy logic based on competencies. As an upgrade of a used model, the authors suggest using of Choquet integral and comparing these two approaches.
There is a recommendation for a system which uses a neuro-fuzzy agent for automatic assessment of each expert and rank criteriums with the management of uncertainty and non-consistency in group decision environment. An approach which ensures identification of the essential characteristics for helping experts is a compound of fuzzy sets and neural networks [13].
The research [14] presents classification techniques in the in-depth analysis of talent assessment. The best results were with the classification algorithm C4.5 where generator tip of decision tree explicitly focuses on relevant attributes.
In [15], one can find the proposal of a framework for in-depth data analysis based on decision tree and associative rules, which generates useful rules for choosing the most appropriate candidate. This framework could find useful laws in data about the relationship between employees and their behavior.

III. Data Science Approaches

In processes of evaluation and selection, an important part is personality profiling since these data have tremendous value in selecting the best candidate. The usual procedure includes standardized test performed by psychologists who have expert knowledge for profiling personality.
However, since the present time is the era of Big data, and since this time characterizes a continually growing quantity of structured and unstructured data as data sources, there is an expectancy for automated analysis and determining personality type.
Every person interaction on the Internet or in real life theoretically gives some source for determining personality type, such as all kind of multimedia data, data from different sensors like cameras, motion sensors, or even any pattern or data as a result of mobile phone usage. For example, the research proposed in [24] shows usage of RNN as a tool for monitoring changes of emotional status based on analysis of the film scenario, and the change of emotional status uses the Big Five personality model.
For the simulation of the process of how a human being is learning, one can use deep neural networks as a computer model. There is identical bias as with human beings – there is a preference of shape comparing the color. Techniques of cognitive psychology can be helpful in better understanding of deep neural networks [18].
However, in focus of this paper are findings in researches which have a spotlight on using natural language processing (NLP) and text data as a source for determining the personality type of text author. For that reason, the spotlight is on using machine learning to discover patterns in a text which are typical for different authors from the perspective of personality traits as the base for models which are in focus, like MBTI.
This section gives an overview of different data science approaches in personality determination.
With the hypothesis that individual writing style and personal characteristics correlate, [20] uses deep learning models in textual date for personality determination. There were used models for unsupervised clustering with the single value decomposition (SVD), recurrent neural network (RNN) with long short-term memory (LSTM), and a bag of words feed-forward neural network. Succession rate for classification of 16 classes was on average 37%. Development of precise model would have a profound impact on business intelligence, analyses of personal compatibility relationships and some other related problems, with the emphasize on additional usage in social sciences [20].
Liu et al. [21] give different approaches to sharing of textual information for RNN, that is relevant in the context of the analysis of information dependent elements, which uses class components analysis.
Some researches propose the use of shallow neural networks for recognition of personality traits of software developer regarding the fragment of program code. The proposed solution uses the Big Five personality type [22].
The research proposed in [26] presents results of personality prediction for the Big Five model on a relatively small dataset of student essays to understand the connection between the language and personality.
Mayuri et al. [16] provide a system for personality prediction, according to the Big Five model. The system uses a multilayer neural network, which works with a group of tweets and does not use the user's profile into account. Also, the work presents the use of Hadoop framework to predict personality traits more individuals at the same time.

Liu et al. [32] use deep learning models and characters to build a hierarchical model C2W2S4PT, to words and sentences representations for personality traits prediction. Character bi-LSTMS construct a word, and word bi-LSTMS construct a sentence. The feed-forward neural network makes a prediction based on Big Five personality traits. The system used PAN 2015 Author Profiling task dataset, collected from Twitter. Annotations in tweets had a purpose of signing age, gender, as well as the Big Five personality traits with support for four languages [34].

Sun et al. [33] present model named 2CLSTM, a bi-LSTMS on word2vec embeddings concatenated with CNN layer for document level representation, to determine a user's personality using text structures. The system evaluates two datasets, long and short texts with better results than in similar experiments.

Mairesse et al. [27] show result in the process of automatic determination of Big Five personality combining communication and text and using ratings of personality as a result of self and observer assessment. Regarding methods, this research uses regression, classification, and ranking models, which gives the best results compared to the baseline and the other two models. However, it uses relatively small dataset, which is as a constraint and area for improving the research. The paper also presents an experiment suggesting better accuracy when using ranking models instead of multi-class classifiers in determining personality type.

Linguistic style is an efficient and appropriate method of determining personality traits [28]. The research examined written language regarding the reliability, factor structure, and validity on datasets of daily diaries, daily writing assignments, and journal abstracts targeting 5-factor measure.

The research proposed in [30] presents using Support Vector Machines on the Five-factor model for determining the personality type of around 3.000 bloggers writing over several months with distinguished results in detecting personality traits of neuroticism and openness, as well that language structure is essential when classifying personality traits.

Bottom-up stratified corpus comparison for examining a collection of e-mail messages created by graduated university students who participated in the experiment of known personality uses [29]. The results show there is little evidence for implicitness of positivity and negativity detecting from observed sources but has found distinguished results in detecting traits of extraversion and neuroticism. The paper confirms that individual differences persist in an e-mail as a data source.

The handwriting text can also be a data source for personality type determination. In research [23], one can see the usage of artificial neural networks for personality type determination from handwriting text.

The research proposed in [19] covers the lexical approach to personality and Big Five type of personality. It also examines a dynamic approach to personality and proposes a neural network model for creating the link between the neurobiology of personality and temper.

Social media is a significant source of data for determining personality type. Publicly available information on the author's Twitter profile can be the source for determining the personality type of author according to Big Five factor model [31]. The results of implementing ZeroR and Gaussian Processes learning algorithms were to within in range from 11% to 18% of their actual value of personality traits. The number of participants was small (n=50) that can be considered as a constraint.

Social networks are generally a gold mine for automated personality prediction, as [25] states for Reddit. The work results in the large-scale dataset (n=9111, 22.9M comments) labeled with self-disclosed MBTI types. The used model combines user activity and linguistic features. Associated results were over 81% in prediction exact or one-off personality type. The feeling dimension was the most difficult to predict (67.2% F1 score).

With using the model which integrates Word Embedding features with Gaussian Process regression, it is possible to reduce the data required for predicting personality traits on Twitter according to the Big Five model, eight times than in state-of-the-art techniques [35].

Park et al. [36] proposed a model for personality determination using an open-vocabulary analysis of the written text of 66.732 Facebook users in addition to results of self -reported Big Five personality traits which are the result of traditional assessment through a questionnaire. The results show valid personality measure and stability over 6-months interval.

Skowron et al. [37] propose an innovative method which integrates different formats, like text, image, and users' metadata features from two social networks: Twitter and Instagram. Proposed outcomes in the prediction of personality traits are better than the base results.

Park and Hovy [38] present the use of Twitter for predicting personality type according to the MBTI model. The source was 1.2M tweets in English for 1.500 participants, and they prove the reliability of prediction for two (introversion and feeling) of four personality dimensions. They combined a logistic regression classifier and word n-grams and meta-features.

Lima and Castro [39] show the research of using Twitter self-reported MBTI types on a relatively small set of 29.200 tweets and a small number of participants (n=146). It proposes mapping MBTI types to Keirsey Temperament Sorter (KTS) and uses several classifiers (SVM, NB). The research results show over 70% accuracy of prediction for two dimensions.

Lin Li et al. [40] shows how the micro-blogging platform can be the data source for determining a user's personality. Regarding classification models in research, it proposes Support Vector Machine (SVM) to distinguish participators regarding the high and low scores by each dimension of the Big Five Inventory. The research uses a micro-blogging platform, Sina Weibo, a popular micro-blogging service in China. The paper shows the accuracy in classification from 84% to 92%. For a regression model, the research uses PaceRegression methods, and for a personality model Big Five model. The Pearson correlation coefficients among presumed and confirmed results were between 0.48 and 0.54. The research proves that data from micro-blogging platforms have the potential for personal traits determination.

Piedboeuf et al. [41] show results in determining personality type from a collection of data on LinkedIn labeled with a DISC or MBTI personality model. Achieved results are in the range 73.7% to 80.5% of precision for the DISC model and in the range 80.7% to 86.2% of precision for the MBTI personality model. Regarding the classifier method, the research uses SVM classifier combined with a feature ranking algorithm and optimization. Also, it uses a Random Forest with AdaBoost and a Naïve Bayes. The results are surprising, because of the professional nature of LinkedIn and expected professional and formal writing style, similar to research results on other social networks where one uses a natural writing style.

IV. Discussion and Conclusion

This paper presents approaches and methods that can improve the conventional method of personality type assessment (e.g., questionnaire survey) towards automatic determining personality traits. Having in mind that small companies cannot afford to have a professional for assessing the traditional approach, that large company face the problem of a large number of candidates and a short time to assess them, that there is always space for human mistake, the importance of an automatic approach to this problem will rise in the future. Also, sometimes we need evaluation for the offline candidate and maybe a candidate who is far away.

Since the majority of the human population spend a considerable amount of time in interaction on social networks, they generate a massive amount of data during their communication. Also, social networks are an essential part of the communication on the Internet, and as a platform, they provide a tremendous opportunity in assessment participant regarding their personality. Social networks have volume significant data sources in different formats, some of it generated in real time. However, social media have just a few publicly available datasets because of privacy issues and labeling costs, which are going to be challenging in the future too. One can also expect new models, not only to have better prediction results but also to mitigate biases in data. However, the proposed results are encouraging, and since the social media are part of everyday life of most of the population, one can expect that it will be the platform for breakthroughs in the field of automated personality prediction.

Because data science heavily depends on the quality and volume of datasets, acquiring proper datasets will be a considerable challenge all that time, especially in attempts to combine results from multiple sources. Additional constraints in acquiring proper datasets are stricter regulation (e.g., General Data Protection Regulation in EU) and technical constraints. One can expect new attempts in combining the synthetic data generation, real data, and data anonymization. That approach will also ask for new models.

Another challenge is to expand work in automatic personality determination in typology-based models, like MBTI, with taking target personality types as classes with compound labels.

Considering that the problem of determining the personality type of text author with Machine Learning naturally includes experts from different professions, breakthroughs are possible by a joint effort of psychologists, linguists, and data scientists.

References

[1] Quenk L.Naomi, "Essentials of Myers-Briggs Type Indicator Assessment," Wiley, New Jersey, 2009.

[2] Cloninger, C. Susan. Theories of Personality: New Jersey: Pearson Prentice Hall, 2004, 4th edition.

[3] J. W. N. E. C. Zahid Hussain, »The use and impact of human resource information systems on human resource management professionals,« Information & Management, Vol. 44, pp. 74-89, 2007.

[4] Harvard Business Review, "Why People Quit Their Jobs," September 2016

[5] Work Institute, http://info.workinstitute.com/retentionreport2018

[6] Society for Human Resouce Management (SHRM), https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/2017-Talent-Acquisition-Benchmarking.pdf

[7] Encyclopedia of Psychological Assessment, Vol 1-2, SAGE Publications, 2003

[8] Caprara, G.V. & Cervone, D, "Personality: Determinants, Dynamics, and Potentials," Cambridge University Press, 2000

[9] K. R. Navneet Bhushan, Strategic Decision Making; Applying the Analytic Hierarchy Process, London: Springer-Verlag, 2004.

[10] P.-C. Chen, »A Fuzzy Multiple Criteria Decision Making Model in Employee Recruitment,« July 2009. [Mrežno]. Available: http://paper.ijcsns.org/07_book/200907/20090716.pdf. (2019-05-15)

[11] N. A. Ruskova, »Decision Support System for Human Resources Appraisal and Selection,« u 2002 First International IEEE Symposium "Intelligent systems," 2002.

[12] E. K. Adem Golec, »A fuzzy model for competency-based employee evaluation and selection,« ScienceDirect/Elsevier, Melikgazi, Kayseri, Turkey, 2007.

[13] H. N. D. R. V. C. Faiyaz Doctor, »A Neuro-Fuzzy Based Agent for Group Decision Support in Applicant Ranking within Human Resources Systems,« in FUZZ-IEEE 2009, Korea, 2009.

[14] A. R. H. a. Z. A. O. Hamidah Jantan, »Data Mining Classification Techniques for Human Talent Forecasting, Knowledge-Oriented Applications in Data Mining,« 2011. http://www.intechopen.com/books/knowledge-oriented-applications-in-data-mining/data-mining-classification-techniques-for-human-talent-forecasting.

[15] F. Ujević I N. BogunovićI," Intelligent Systems at the service of Mankind, Wilfried Elmenreich, J. A. Tenreiro Machado, and Imre J. Rudas, 2003, p. 457.

[16] Mayuri Pundlik Kalghatgi, Manjula Ramannavar, Dr. Nandini S. Sidnal: "A Neural Network Approach to Personality Prediction based on the Big-Five Model," 2015, URL: http://www.ijirae.com/volumes/Vol2/iss8/09.AUAE10095.pdf (2018-02-07)

[17] C. K. a. B. Oztaysi, »Personnel selection using type-2 fuzzy aph method,« The Business&Management Review, Volume 4 Number 1, 2013.

[18] Samuel Ritter, David G.T. Barret, Adam Santoro, Matt M. Botvinick: "Cognitive psychology for Deep Neural Networks: A shape bias case study," 2017, URL: https://arxiv.org/pdf/1706.08606.pdf (2018-02-07)

[19] Read, S. J. et al. A Neural Network Model of the Structure and Dynamics of Human Personality. Psychological Review 117, 61–92 (2010)

[20] Anthony Ma, Gus Liu: "Neural Networks in Predicting Myers Briggs Personality Type from Writing Style," 2017 URL: https://web.stanford.edu/class/cs224n/reports/2736946.pdf (2018-02-07), unpublished.

[21] Pengfei Liu, Xipeng Qiu, Xuanjing Huang: "Recurrent Neural network for Text Classification with Multi-Task learning," 2016 URL: https://arxiv.org/abs/1605.05101 (2018-02-07)

[22] Yerai Doval, Carlos Gomes-Rodriguez, Jesus Vilares: "Shallow recurrent NN for personality recognition in the source code," 2016 URL: http://ceur-ws.org/Vol-1737/T1-6.pdf (2018-02-07)

[23] Champa H N, Dr. K RK AnandaKumar: "Artificial-Neural-Network-for-Human-Behavior-Prediction-through-Handwriting-Analysis," 2010, International Journal Of Computer Applications (0975-8887) Volume 2-No.2 May 2010

[24] Takumi Ichimura, Kosuke Tanabe, Toshiyuki Yamashita: "An Adaptive Learning Method of Personality Trait-Based Mood in Mental State Transition Network by Recurrent Neural Network," 2014, URL: http://ieeexplore.ieee.org/document/6988081/ (2018-02-07)

[25] Matej Gjurković and Jan Šnajder, 2018. "Reddit: A gold mine for personality prediction." The Second Workshop on Computational

Modeling of People Opinions, Personality, and Emotions in Social Media, Proceedings, pages 87-97

[26] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, “Lexical predictors of personality type,” in Proceedings of the Joint Annual Meeting of the Interface and the Classification Society, 2005, pp. 1–16

[27] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using Linguistic Cues for the Automatic Recognition of Personality in conversation and text,” Journal of AI research, vol. 30, pp. 457–500, 2007 https://www.aaai.org/Papers/JAIR/Vol30/JAIR-3012.pdf

[28] J. W. Pennebaker and L. A. King, “Linguistic styles: Language use as an individual difference,” Journal of Personality and Social Psychology, vol. 77, no. 6, p. 1296, 1999

[29] J. Oberlander and A. J. Gill, “Language with character: A stratified corpus comparison of individual differences in e-mail communication,” Discourse Processes, vol. 42, no. 3, pp. 239–270, 2006

[30] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, “Large scale personality classification of bloggers,” in Affective Computing and Intelligent Interaction. Springer, 2011, pp. 568–577

[31] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, “Predicting personality from Twitter,” in Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE International Conference on Social Computing (SocialCom), 2011, pp. 149–156

[32] F. Liu, J. Perez, and S. Nowson, “A recurrent and compositional model for personality trait recognition from short texts,” in Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES). The COLING 2016 Organizing Committee, 2016, pp. 20–29. [Online]. Available: http://www.aclweb.org/anthology/W16-4303

[33] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, “Who Am I? Personality Detection Based on Deep Learning for Texts,” in 2018 IEEE International Conference on Communications (ICC), May 2018, pp. 1–6.

[34] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, “Overview of the 3rd author profiling task at PAN2015,” in CLEF 2015 labs and workshops, notebook papers, CEUR Workshop Proceedings, vol. 1391, 2015

[35] P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha, “25 tweets to know you: A new model to predict personality with social media,” in Proceedings of the Eleventh International AAAI Conference on Web and Social Media, 2017, pp. 472–475.

[36] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, D. J. Stillwell, M. Kosinski, L. H. Ungar, and M. E. Seligman, “Automatic personality assessment through social media language,” Journal of Personality and Social Psychology, vol. 108, pp. 934–952, 2014

[37] M. Skowron, M. Tkalcic, B. Ferwerda, and M. Schedl, “Fusing social media cues: Personality prediction from Twitter and Instagram,” in Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016, pp. 107–108.

[38] B. Plank and D. Hovy, “Personality traits on Twitter –or– how to get 1,500 personality tests in a week,” in Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), 2015, pp. 92–98. [Online]. Available: http://www.aclweb.org/anthology/W15-2913

[39] A. C. E. Lima and L. N. de Castro, “Predicting temperament from twitter data,” in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, 2016, pp. 599–604.

[40] Lin Li, Ang Li, Bibo Hao, Zengda Guan, Tingshao Zhu: "Predicting Active Users’ Personality Based on Micro-Blogging Behaviors," 2014, URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084997 (2018-02-07)

[41] F. Piedboeuf, P. Langlais, L. Bourg, “Personality Extraction Through LinkedIn,” Advances in Artificial Intelligence, 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019 Kingston, ON, Canada, May 28-31, 2019, Proceedings, Springer 2019