

Enhancing Human Resources through Data Science: a Case in Recruiting

Paolo Frazzetto^{1,2}, Muhammad Uzair Ul Haq^{1,2} and Alessandro Sperduti¹

¹*Department of Mathematics “Tullio Levi-Civita”, University of Padova, 35121 Padua, Italy*

²*Amajor S.r.l. SB, via Noventana 192, 35027 Noventa Padovana, Italy*

Abstract

The role of Human Resources (HR) in the recruitment process is undergoing great technological changes with the increasing use of web-based job portals for candidate selection. While these portals have improved efficiency for applicants, they have also presented challenges to recruiters in managing the large volume of applications received. Automated systems that use machine learning have been proposed to address this, but the lack of publicly available annotated datasets hinders progress in this field. To overcome this limitation, we introduce a novel dataset of Italian CV embeddings encoded with binary targets, making it publicly accessible for research purposes. The study further explores the performances of data science techniques on the proposed dataset and the application of a Network Science perspective.

Keywords

HR, Personnel selection, CV Dataset, AutoML, Graph Neural Network, Machine Learning

1. Introduction

Human Resources (HR) play a central role in any business—finding and hiring the best candidates is of paramount importance to ensure the long-term growth, productivity, and competitiveness of an organization [1].

The process of identifying and selecting suitable candidates for a job has traditionally relied on manual assessment, interviews, and subjective judgments. However, in recent years there has been increased interest in recruiting through web-based job portals [2]. Online job portals allow Human Resource Management (HRM) to target a larger audience and boost candidates’ reach-out. With the help of these platforms, candidates can easily upload their data, supporting documents, such as Curriculum Vitae (CV) or video presentations, or fill in assessment questionnaires [3].


By using such systems, the Human Resources (HR) department can receive a large number of applications, even for a single job posting. On the one hand, these systems have made the job application process efficient for candidates. Still, on the other hand, it has made the screening process time-consuming and labor-intensive for recruiters. Therefore, an automated system is desired. Recently, there have been many efforts to employ the use of machine learning to solve such tasks [4, 5, 6]. However, an annotated dataset is required to use the full potential


ITADATA2023: The 2nd Italian Conference on Big Data and Data Science, September 11–13, 2023, Naples, Italy

✉ paolo.frazzetto@phd.unipd.it (P. Frazzetto); muhammaduzair.ulhaq@phd.unipd.it (M. U. U. Haq);

alessandro.sperduti@unipd.it (A. Sperduti)

ORCID 0000-0002-3227-0019 (P. Frazzetto); 0000-0001-9660-8982 (M. U. U. Haq); 0000-0002-8686-850X (A. Sperduti)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of Machine Learning. Due to privacy concerns, most companies do not tend to release their dataset to the public, thus hampering the basic research in this field.

In this paper, we analyze and release a novel dataset of CV embeddings. The CVs are in the Italian language, embedded using a multilingual Large Language Model (LLM), and their targets are encoded in a binary fashion. In this way, an alternative solution is proposed to make the dataset publicly available. The embeddings can easily be used for data analysis and building machine learning models. Each candidate is labeled by his progress in the corresponding selection. We also explore how Data Science techniques can enhance the candidate selection process, improve decision-making, and ultimately contribute to the success of organizations. We further investigate how we can exploit questionnaire data from a Network Science perspective, testing whether similar questionnaires result in similar candidates' outcomes. Our ultimate goal is to contribute to this field's growing body of knowledge and guide HR professionals in adopting data-driven approaches to optimize their talent acquisition strategies.

Contributions of the paper: (i) we propose a novel publicly available HR recruitment dataset obtained by collecting job candidates' data from real-world cases; (ii) we present results on the proposed dataset by many strong baseline models, including Neural Networks (NN) and Graph Neural Networks (GNN).

2. Related Work

Nowadays, automated extraction of information from CVs, job postings, and related HR documents has caught the attention of many researchers and companies [5, 6, 7]. However, the lack of publicly available datasets bottlenecks most of the progress.

Despite the scarcity of datasets, there have been some efforts to extract information from resumes and job postings with the aim of boosting HR performance. For example, [8] proposed the use of an average word embedding model for CV retrieval based on the job description. The embeddings of CVs are trained from scratch and combined with the pre-trained word2vec embeddings using a hybrid embedding model. The job postings are also embedded using a pre-trained word2vec model, and cosine similarity is used as a measure to find relevance between the CV and job description. However, to the best of our knowledge, the dataset used in the research is not publicly available. Jiang et al. [9] proposed the use of machine learning to match candidates with job postings on online platforms. The authors used real-case data collected by the recruitment team between January 2019 and October 2019. They collected about 13K jobs with 580K candidates and 1.3 million resumes. However, the authors claim the dataset to be sensitive and do not release it publicly. Yao et al. [10] proposed a method to quantify the matching between candidate qualifications and the job requirements of a position. They model the task with distantly supervised skill extraction to identify the skill entities from job postings and resumes using skill entity dictionaries. The relevance between a resume and a job description is measured according to the matching score of the skill entities. The dataset used in the research contains 21K job postings and 86K resumes provided by a high-tech company. So far, the authors have not released the publicly available dataset.

Zhang et al. [11] released a novel dataset for skill extraction on English job postings called SKILLSPAN. The authors also outlined the annotation guidelines created by domain experts to

annotate hard and soft skills in job postings. The dataset consists of 14.5K sentences, of which 12.5K are annotated. The dataset is divided into three categories; BIG, HOUSE, and TECH. The authors only released the HOUSE and TECH categories of the dataset to the public. However, the dataset is limited to job postings, and only skill entities are annotated.

Given the limited availability of publicly available datasets, in this study, we address the problem at hand and release a novel dataset of CV embeddings. More details of the dataset are explained in Section 4.

3. Data Collection

The gathering of reliable data in HR recruitment is a complex task that goes beyond the realm of academia, as it requires external support from the industry. While academic research provides valuable insights and theoretical frameworks, it often falls short of capturing the intricacies and practicalities of the recruitment process in real-world settings. Additionally, HR recruitment involves gathering a wide range of data, including resumes, application forms, psychometric assessments, interviews, and performance evaluations. This data collection process requires collaboration with organizations willing to share their recruitment data and provide access to their internal information systems.

This work has been made possible by the partnership and support of Amajor S.r.l SB¹, an Italian business development consulting firm. Its main activities concern proprietary consulting methods to guide small- and medium-sized enterprises, improve their business model through the entrepreneurs' values, and find new candidates that best fit within the client's organization.

This partnership enabled us to obtain authentic and diverse datasets, allowing us to analyze and develop models that closely mirror the challenges and complexities faced by HR practitioners. This collaborative approach ensures that our research findings are relevant, applicable, and aligned with the practical needs of the industry, ultimately leading to more effective and impactful HR recruitment strategies.

3.1. Privacy Regulations

Privacy is a paramount concern in the field of HR, especially when dealing with the sensitive personal data of job applicants. This sensitive data falls under specific legal frameworks, most notably the General Data Protection Regulation (GDPR) in the European Union [12]. This regulation mandates that organizations handle personal data with care, ensuring its confidentiality, security, and lawful processing.

Furthermore, with the increasing integration of AI technologies in HR processes, additional privacy concerns arise: such as the potential for unintended biases, discriminatory practices, and unauthorized access to personal information. These and broader applications are being regulated in the EU by the so-called AI Act [13] which, once approved, will constitute the world's first governmental regulation on Artificial Intelligence. Specifically, AI applications in HR are classified as *High-Risk AI Systems* as they are mentioned in Annex III.4.(a):

¹Corporate Website: <https://www.amajorsb.com/en/>

“AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;”

Therefore, HR professionals must navigate these specific legal requirements by implementing robust data protection measures, obtaining informed consent from candidates, and ensuring secure storage and transmission of personal information.

In the scope of this work, we have taken extensive measures to handle data in a compliant manner. Prior to data collection and questionnaire administration, candidates have been duly informed about the purpose and scope of data processing, providing their informed consent. In addition, we ensure anonymity by removing any personally identifiable information from the dataset and releasing the CV embeddings rather than their original form. We analyzed the data in its aggregate structure without any additional bias. Besides, we adhere to data retention policies and will promptly delete the original records once they are no longer necessary for the intended purposes. In this fashion, we aim to maintain the confidentiality and trust of the candidates' personal data, along with promoting open science and productive discussion in the field.

4. Dataset Description

The data have been collected from real-case candidates' applications for 195 different job postings for vacancies, mostly in North-East Italy. The positions vary in terms of seniority, role, business area, and corporate size. The time frame spans from 01-01-2021 to 31-05-2023. From a starting pool of more than 13.000 applications, we filtered out those candidates who did not fill in the privacy consent for this research, those who did not complete the assessment questionnaire, and those who did not submit their CV.

This would be the first release of such a dataset, and we plan to provide future releases with more entries, features, and meta-data. In fact, we intend to extend our study by incorporating meta-data from the job description and the entire selection process. The dataset is available at the repository [Research Data Unipd](#).

4.1. Classes Identification

The job selection process typically involves several steps that candidates must navigate. In our scenario, candidates are first required to submit their application, which includes their CV and other basic data, and fill in the assessment questionnaire. After the initial screening by HR recruiters, candidates are invited to participate in one or more interviews, which can be conducted in various formats, such as video interviews or in-person meetings. From the outcome of the interviews, only a handful of candidates are shortlisted and presented to the future employer. In the final stage, the employer selects one candidate, negotiates with him the job offers, reviews the employment contracts, and completes the necessary paperwork before officially joining the organization.

In a Machine Learning framework, we can consider each of the previous states as a label, thus modeling the process as a multi-class classification problem. Alternatively, since the process is

consequential, one could model it as a regression task over an interval. Nonetheless, for this first analysis and release of the dataset, we opted to consider binary labels. We realized that in real-case applications, there are many deviations from this standard process; besides, the data recorded in the information systems may be miss-classified or missing for some candidates and stages. Therefore, we focus on those candidates who completed all stages up to the first interview—an HR specialist has checked their CV, questionnaire, and interview and made the decision to bring them further in the selection (positive labels) or deemed that they were not the best-suited candidates for that role (negative labels). These steps left out $N = 2,647$ valid candidates, of which 1,674 (63%) with positive labels and 973 (37%) with negative ones. Meanwhile, we are currently working on cleaning and pre-processing the datasets to get more reliable data points.

4.2. CV Embeddings

State-of-the-art Large Language Models (LLMs) have succeeded in various natural language processing tasks. These models can be pre-trained on large corpora to capture contextual information of words in a text. CVs are unstructured documents that consist of long textual information. In this study, we use an XLM-RoBERTa-Longformer [14], a multilingual model with an input size of 4096 tokens. The multilingual characteristic allows us to capture information in different languages, whereas the larger input size enables processing long documents.

The documents are preprocessed by removing stopwords, extra spaces, and special characters. These documents are tokenized and then passed through the pre-trained XLM-RoBERTa-Longformer to extract the word embeddings of all the tokens. Each token is represented by a 768-dimensional feature vector. Therefore, a document consisting of N tokens returns a $N \times 768$ dimensional feature matrix. Processing such large matrices is computationally expensive; therefore, we average the embedding vector of all the tokens in the document, resulting in a 768-dimensional feature vector representing the document in an embedding space.

4.3. Questionnaire Data

Personality and behavior assessments through questionnaires are one valuable tool in HR selection processes and organizational psychology [3]. These assessments aim to gain insight into candidates' individual traits, characteristics, and behavioral tendencies, providing a deeper understanding of their fit within the organization and their job role. Some questionnaires are designed specifically for personality assessment, others for measuring some ability or behavior. There exists a plethora of different kinds of tests, with different validity and usage among HR practitioners [15]. Nevertheless, the quantitative nature of questionnaire data, along with its ease of collection, allows for rigorous statistical analysis and enables standardized evaluation across candidates, promoting fairness and consistency in the selection process.

We collected questionnaire data following Amajor's business model. The tool used for the candidates' assessment is the so-called A+ Questionnaire: a set of 242 questions with 3-scale Likert-type answers (yes/maybe/no or similar) covering various aspects of one's behavior, habits, and personality, developed by the company team after working alongside more than 120 clients over a period of 3 years [16]. The answers are grouped and processed following a proprietary

factor model that gives an estimate of one’s hidden traits; however, its analysis and discussion go beyond the scope of this work. In the following section, we describe a novel and general approach to exploit Likert-type questionnaire data to find patterns among respondents.

4.3.1. From Questionnaires to Graph

Likert-type scales are widely employed in academic and industrial settings to capture human facets due to their user-friendly nature, simplicity of development, and ease of administration [17]. It enables respondents to answer questions in a closed-form way, picking only one value on an ordered scale according to some sort of preference or agreement. Due to the fact that the perceived distance between two consecutive items cannot be defined or presumed equal [18], such a scale cannot be analyzed by classical statistical methods defined on a metric space or parametric tests but requires specific modeling and assumptions [19].

In order to link candidates that give similar answers, we resort to Network Science. In fact, networks, also called graphs, enable more expressive data structure and occur in many fields of science and engineering [20]. However, translating tabular data to graphs is not trivial [21, 22] as it requires domain knowledge and heuristics to define the nodes and their relationships.

Our approach to tackling these issues is straightforward and takes advantage of the specific structure of Likert-type data. Given any Likert scale Questionnaire $\mathcal{Q} = \{q_1, q_2, \dots, q_n\}$ made up of n questions, each possible answer a_i takes value in an ordered set that w.l.o.g. we can define as $\mathcal{A} = \{1, 2, \dots, L\} \subset \mathbb{N}^n$, so $a_i \in \mathcal{A}$, $\forall i = 1, \dots, n$. The order relation depends on each q_i , and we assume that it is universal, i.e., the questionnaire is well-written, and each question is understood by all respondents. In this way, each completed questionnaire can be formulated as a specific collection of all the possible answers $\mathbf{a} = \{a_1, \dots, a_n\}$, $\forall a_i \in \mathcal{A}$ and a respondent can be described as a function $r : \mathcal{Q} \rightarrow \mathcal{A} \times \dots \times \mathcal{A} = \mathcal{A}^n$, $r(\mathcal{Q}) = \mathbf{a}$. We desire to link candidates/respondents that provide similar answers, thus having similar behaviors and personalities, without resorting to the hidden variables given by factor analysis. Next, we have to define a distance function $d(\mathbf{u}, \mathbf{v})$ between two responses \mathbf{u}, \mathbf{v} . Our desiderata are that respondents who give the exact same answers will be closer, whereas when the answers are on the opposite side of the scale, the distances should be greater. Additionally, we want to avoid the Euclidean metric, since it scales quadratically with L , but Likert scales are perceived as linear [18]. Therefore, the ideal candidate is the Manhattan distance or l_1 norm:

$$d_M(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n |u_i - v_i| \quad u_i, v_i \in \mathcal{A}. \quad (1)$$

We brought this idea further by considering that Likert-type answers are usually contrasting, i.e., one end of the scale is the opposite of the other, with all the ranges in between. Therefore, in the case of an odd number of choices L , a middle value is perceived as a neutral or indefinite answer [23]. We want to emphasize this contrast and penalize the neutral answers, as they provide little insight into the analysis. For these reasons, we center our answer set in zero $\tilde{\mathcal{A}} = \{-\lfloor L/2 \rfloor, \dots, 0, \dots, \lfloor L/2 \rfloor\}$ and we exploit this symmetry with a redefined Bray-Curtis similarity [24]:

$$d_{BC}(\mathbf{u}, \mathbf{v}) = 1 - \frac{\sum_i |u_i - v_i|}{\sum_i |u_i + v_i|}. \quad (2)$$

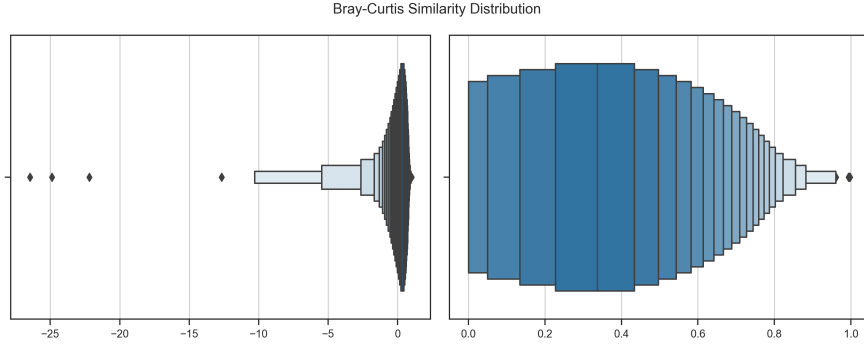


Figure 1: Boxenplot distribution of the Bray-Curtis similarity for all N^2 questionnaire pairs; before (left) and after (right) clipping the negative values.

Notice that this measure of similarity has the desired properties, being normalized to 1 when the answers of two questionnaires are exactly the same. On the other hand, it diverges to $-\infty$ when they are always at the opposite. In practice, $d_{BC}(\mathbf{u}, \mathbf{v}) \leq 0 \iff \sum_i |u_i - v_i| \geq \sum_i |u_i + v_i|$ and the latter holds when the majority of answers have opposite signs, hence meaning.

In our specific business case, we have $L = 3$ and thus $\mathcal{A} = \{-1, 0, 1\}$, where 1 stands for the positive/affirmative answer, 0 is maybe/neutral, and -1 is no/negative. We then computed the pairwise similarity of Eq. (2) for all pairs. These values directly translate into a graph in which each node is a candidate, who is connected to all other candidates by means of a weighted adjacency matrix A , whose entries are consequently defined as $A_{uv} = d_{BC}(u, v) = A_{vu}$. In this way, we obtain a fully connected graph with an order of $N^2 \simeq 3.5 \times 10^6$ links. We also apply two heuristics to reduce its complexity and keep only the most significant connections. First, we set the negative values to zero, enforcing no similarity between such different questionnaires. This results in the removal of 1.47×10^5 links. The corresponding boxenplot distribution is shown in Figure 1, noticing that we retain a homogeneous distribution of similarities along a right-tail of candidates with almost identical answers. Secondly, our aim is to drop the links with a weight close to zero, starting from the lowest values. Therefore, we apply edge percolation [25] and heuristically stop when the largest component has 95% of the total nodes. As shown in Figure 2, an additional 1.1×10^5 edges can be removed. The corresponding threshold of the similarity $d_{BC}(u, v)$ is 0.07, so all the remaining links connect candidates with a similarity greater than this threshold. The basic statistics on the obtained graph are reported in Table 1.

Such a process allowed us to find a reasonable graph of candidates based on their responses on a Likert-type questionnaire. Our research question is to test whether such an approach can improve the identification of patterns and the prediction of the class of new candidates, given that they provided similar answers to other labeled candidates.

5. Candidates Classification

This section explores how Machine Learning can be leveraged to perform candidate classification based on this novel dataset. Two different approaches have been investigated—relying on

Table 1
Basic Properties of the Candidates Graph

Property	#Nodes	#Links	Avg. Degree	Avg. Clustering Coeff.	Avg. Path Length	Graph Diameter	Graph Density
Value	2647	518994	392	0.615	2.141	8	0.148

unstructured or structured data.

The first approach focused on tabular data analysis, where traditional machine learning algorithms were applied to extract insights and patterns from unstructured data. This pipeline usually involves feature engineering, model selection, architecture search, hyper-parameter optimizations, and training on the tabular dataset to make predictions and classifications. Each of these steps may be challenging on its own; therefore, we resorted to off-the-shelf AutoML tools to automate this procedure. In particular, we exploited AutoGluon [26] for its simplicity and the availability of Neural Networks among its models.

In addition to the tabular approach, we also explored the use of Graph Neural Networks (GNNs) [27, 28] to analyze the dataset’s graph structures as described in Section 4.3.1. By representing the data as a graph, we leveraged GNNs to capture the relationships and dependencies that emerge from questionnaire data among the entities. The GNN model enables us to learn from both the nodes’ attributes (i.e., the CV embeddings) and the relational information present in the graph, thereby capturing complex patterns and interactions that might be missed by traditional tabular approaches. Contrary to tabular or multi-modal data, AutoML tools for GNNs are still in their infancy and are under active development [29]. However, we adopted our graph for the AutoGL framework [30], which enables us to test some of the most common GNN layers for the node classification task.

By employing these two complementary approaches, we aimed to gain a comprehensive

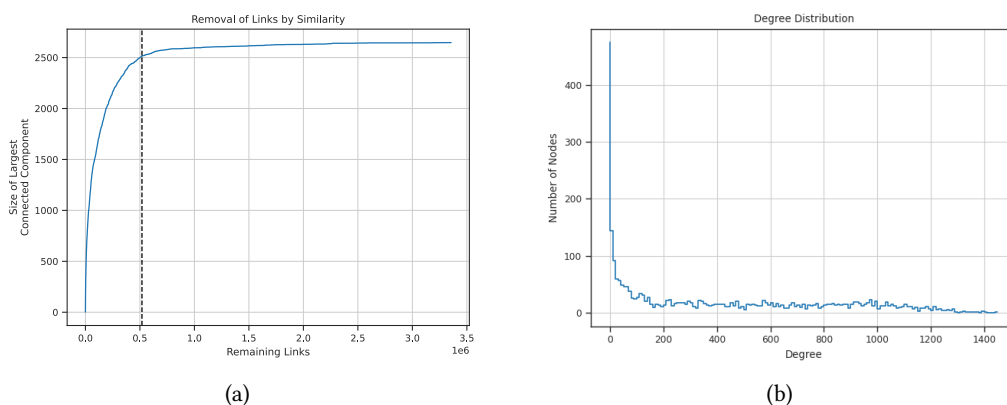


Figure 2: (a): Size of the largest connected component by removal of the links with ascending similarity. The dotted line indicates the point where the largest component has 95% of the original nodes; thus, we retain all the connections up to that point. (b): Degree distribution of the resulting graph.

Table 2

Experimental results on the test set for different models and data structures.

Data	Model	Accuracy [%]	Std. Dev.
Baseline	Class Prior Classifier	63.24	-
Tabular (CV Embeddings)	NN	63.40	2.16
	RandomForest	61.51	1.79
	CatBoost	64.53	1.51
Tabular (Questionnaire)	NN	75.78	2.54
	RandomForest	76.99	1.42
	CatBoost	77.16	1.20
Tabular (CV Emb. + Qst.)	NN	76.45	2.44
	RandomForest	75.27	1.09
	CatBoost	76.91	1.46
Graph (CV Emb.)	GraphSAGE	77.35	1.77
	GAT	76.60	1.96
	GCN	74.33	2.77

understanding of the dataset and extract meaningful insights from different perspectives. All the experiments have been conducted in the same environment and by employing open-source libraries. We tested different models for each scenario, with 10-fold cross-validation on a [80, 10, 10] train/validation/test split. We disabled any other feature selection/engineering techniques, as we already employ features extracted from raw HR data, and the models’ comparisons would be unfair. For the same reason, we turned off bagging and multi-layer stack ensembling useful to boost predictive accuracy [31].

5.1. Results

The experimental results are reported in Table 2. Our baseline model is a naive class prior probability classifier that always predicts the most common class (the “positive” candidates) with an expected accuracy of 63.24%. We considered RandomForest [32] and CatBoost [33] since they have been proven to be fast to train and effective on tabular data in many domains. Concerning the GNNs, we selected the modules GraphSAGE [34], GAT [35], and CGN [36]. The Neural Networks and GNNs were trained with the default hyperparameters and architecture spaces, therefore, their performances could improve with more extensive model selections.

CV embeddings alone are substantially equivalent to the class prior classifier. This suggests that CV embeddings should be improved, and the current general-purpose LLM are unable to grasp essential information without any domain knowledge or fine-tuning. Considering that the Questionnaire is an essential element in the considered HR selection process, the fact that the answers alone are a better predictor is unsurprising. CatBoost performs the best, being designed for categorical data such as Likert-type questionnaires. Adding the embeddings along with questionnaire data does not lead to improvements but rather results in slightly degraded performances. Therefore, the CV embeddings are not as informative, and also considering their

high dimensionality, they deteriorate classification.

In spite of that, the graph topology we proposed turns out to be valuable for classification, performing slightly better than the corresponding tabular dataset of embeddings plus answers to the questionnaire. It seems that our rationale for linking Likert-type data is effectively linking similar candidates in a meaningful fashion.

6. Conclusions

In conclusion, we collected, processed, and released an HR recruitment dataset. We employed LLM to both preserve anonymity and study the current boundaries of such models in this domain. We analyzed this dataset as a data-driven process, modeling it as a binary classification task. Standard Machine Learning techniques proved effective when combined with assessment questionnaire data, and we demonstrated a manner to convert Likert-type data to graphs while preserving their intrinsic patterns and relations.

In the future, we plan to improve CV embeddings in order to outperform baselines. Additionally, we are already working on collecting more data and meta-data to enlarge the current dataset. Given these promising results, we plan to investigate how to translate questionnaire data into graphs by analyzing different metrics, pruning techniques, and its validity on other questionnaires. Ultimately, we wish to contribute to the advancement of knowledge in this area and provide guidance to HR professionals in adopting data-driven approaches for optimizing talent acquisition strategies.

Acknowledgments

The authors would like to thank the HR recruiters and employees of Amajor for making this research possible.

References

- [1] R. A. Noe, J. R. Hollenbeck, B. A. Gerhart, P. M. Wright, *Fundamentals of human resource management*, McGraw-Hill Education New York, NY, 2016.
- [2] S. Strohmeier, Concepts of e-hrm consequences: a categorisation, review and suggestion, *The International Journal of Human Resource Management* 20 (2009) 528–543.
- [3] R. Bailey, Hr applications of psychometrics, *Psychometric Testing: Critical Perspectives* (2017) 85–111.
- [4] S. Strohmeier, *Handbook of Research on Artificial Intelligence in Human Resource Management*, Edward Elgar Publishing, 2022.
- [5] C. Bizer, R. Heese, M. Mochól, R. Oldakowski, R. Tolksdorf, R. Eckstein, The impact of semantic web technologies on job recruitment processes, in: *Wirtschaftsinformatik*, 2005.
- [6] K. Yu, G. Guan, M. Zhou, Resume information extraction with cascaded hybrid model (2005).

- [7] X. Yi, J. Allan, W. B. Croft, Matching resumes and jobs based on relevance models, in: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007.
- [8] F. C. Fernández-Reyes, S. Shinde, Cv retrieval system based on job description matching using hybrid word embeddings, *Computer Speech & Language* 56 (2019) 73–79. URL: <https://www.sciencedirect.com/science/article/pii/S0885230817302851>. doi:<https://doi.org/10.1016/j.csl.2019.01.003>.
- [9] J. Jiang, S. Ye, W. Wang, J. Xu, X. Luo, Learning effective representations for person-job fit by feature fusion, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2549–2556. URL: <https://doi.org/10.1145/3340531.3412717>. doi:10.1145/3340531.3412717.
- [10] K. Yao, J. Zhang, C. Qin, P. Wang, H. Zhu, H. Xiong, Knowledge enhanced person-job fit for talent recruitment, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), 2022, pp. 3467–3480. doi:10.1109/ICDE53745.2022.00325.
- [11] M. Zhang, K. N. Jensen, S. D. Sonniks, B. Plank, Skillspan: Hard and soft skill extraction from english job postings, in: North American Chapter of the Association for Computational Linguistics, 2022.
- [12] Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, 2016. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, GDPR consolidated version.
- [13] Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, European Commission, 2021. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>, cOM(2021) 206 final.
- [14] Hugging Face, Longformer, <https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096>, Accessed: 2023-06-29.
- [15] A. Furnham, HR professionals' beliefs about, and knowledge of, assessment techniques and psychometric tests, *International Journal of Selection and Assessment* 16 (2008) 300–305.
- [16] E. Peronato, F. Fabris, N. D'Agnolo, R. D'Orazio, Entrepreneurial values as a key for csr in smes, Presented at the XXXIII ISPIM, Copenhagen, 2022, 2022. URL: <https://www.conferencesubmissions.com/ispim/copenhagen2022/index.html>.
- [17] A. Joshi, S. Kale, S. Chandel, D. K. Pal, Likert scale: Explored and explained, *British journal of applied science & technology* 7 (2015) 396.
- [18] J. Munshi, A method for constructing likert scales, 2014.
- [19] M. Disegna, N. Biasetton, E. Barzizza, L. Salmaso, A new adaptive membership function with CUB uncertainty with application to cluster analysis of Likert-Type data, Available at SSRN 4115553 (2022).
- [20] A.-L. Barabási, Network science, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (2013) 20120375.
- [21] J. Liu, Y. Chabot, R. Troncy, V.-P. Huynh, T. Labbé, P. Monnin, From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods, *Journal*

- of Web Semantics 76 (2023) 100761. URL: <https://www.sciencedirect.com/science/article/pii/S1570826822000452>. doi:<https://doi.org/10.1016/j.websem.2022.100761>.
- [22] K. Zhou, Z. Liu, R. Chen, L. Li, S.-H. Choi, X. Hu, Table2graph: Transforming tabular data to unified weighted graph, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 2420–2426. URL: <https://doi.org/10.24963/ijcai.2022/336>. doi:10.24963/ijcai.2022/336, main Track.
 - [23] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, S. L. Young, Best practices for developing and validating scales for health, social, and behavioral research: a primer, *Frontiers in public health* 6 (2018) 149.
 - [24] J. R. Bray, J. T. Curtis, An ordination of the upland forest communities of southern wisconsin, *Ecological monographs* 27 (1957) 326–349.
 - [25] M. E. J. Newman, R. M. Ziff, Fast monte carlo algorithm for site or bond percolation, *Phys. Rev. E* 64 (2001) 016706. URL: <https://link.aps.org/doi/10.1103/PhysRevE.64.016706>. doi:10.1103/PhysRevE.64.016706.
 - [26] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, arXiv preprint arXiv:2003.06505 (2020).
 - [27] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems* 32 (2021) 4–24. doi:10.1109/TNNLS.2020.2978386.
 - [28] A. Sperduti, A. Starita, Supervised neural networks for the classification of structures, *IEEE Transactions on Neural Networks* 8 (1997) 714–735. doi:10.1109/72.572108.
 - [29] K. Cao, J. You, J. Liu, J. Leskovec, Autotransfer: Automl with knowledge transfer – an application to graph neural networks, 2023. arXiv:2303.07669.
 - [30] C. Guan, Z. Zhang, H. Li, H. Chang, Z. Zhang, Y. Qin, J. Jiang, X. Wang, W. Zhu, AutoGL: A library for automated graph learning, in: ICLR 2021 Workshop on Geometrical and Topological Representation Learning, 2021. URL: <https://openreview.net/forum?id=0yHwpLeInDn>.
 - [31] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble selection from libraries of models, in: Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 18. URL: <https://doi.org/10.1145/1015330.1015432>. doi:10.1145/1015330.1015432.
 - [32] L. Breiman, Random forests, *Machine learning* 45 (2001) 5–32.
 - [33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, *Advances in neural information processing systems* 31 (2018).
 - [34] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, *Advances in neural information processing systems* 30 (2017).
 - [35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903 (2017).
 - [36] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2017. arXiv:1609.02907.