Introducing the Djinni Recruitment Dataset: A Corpus of Anonymized CVs and Job Postings

Nazarii Drushchak, Mariana Romanyshyn
Ukrainian Catholic University, Grammarly
Lviv, Ukraine, Kyiv, Ukraine
drushchak.pn@ucu.edu.ua, mariana.romanyshyn@grammarly.com

Abstract

This paper introduces the Djinni Recruitment Dataset, a large-scale open-source corpus of candidate profiles and job descriptions. With over 150,000 jobs and 230,000 candidates, the dataset includes samples in English and Ukrainian, thereby facilitating advancements in the recruitment domain of natural language processing (NLP) for both languages. It is one of the first open-source corpora in the recruitment domain, opening up new opportunities for AI-driven recruitment technologies and related fields. Notably, the dataset is accessible under the MIT license, encouraging widespread adoption for both scientific research and commercial projects.

Keywords: Recruitment Dataset, Open-Source Corpus, Natural Language Processing (NLP)

1. Introduction

This paper introduces the Djinni Recruitment Dataset[1], a unique asset to NLP research in the recruitment domain, where open data is exceptionally limited. The corpus addresses the need for diverse publicly available datasets, which are particularly important in the age of transformers and large language models, especially for low-resource languages such as Ukrainian.

The data for the corpus was provided by Djinni[2], an IT job platform that hosts job listings and anonymized user profiles similar to resumes. Djinni's database is distinguished by its bilingual nature, encompassing both Ukrainian and English languages. The company generously shared with us the data covering a period from 2020 to 2023.

The Djinni Recruitment Dataset opens avenues for various research opportunities. Based on this data, we can analyze the impact of global events on hiring trends and develop recommendation systems tailored to the recruitment domain. The dataset also holds promise for addressing ethical concerns in hiring systems. A corpus of anonymized candidate profiles used for training may help increase fairness in tools like Amazon's AI recruiting tool (Dastin, 2018), which was trained on predominantly male CVs and subsequently exemplified gender bias. The dataset will help promote Responsible AI practices and contribute to the broader discourse on improving the recruitment process.

In this paper, we describe the Djinni Recruitment Dataset and its application. Section 2 reviews related work. Section 3 presents a thorough dataset overview, including source, collection, preprocessing, and characteristics. Section 4 identifies recoverable protected attributes from anonymous CVs. Section 5 discusses the intended use of the dataset in industry and academia. Section 6 addresses the challenges and limitations of the Djinni Recruitment Dataset. Section 7 summarizes the findings and suggests future research directions. Section 8 considers ethical aspects, focusing on privacy and anonymization.

2. Related Work

The exploration of linguistic resources for the Ukrainian language and job-related datasets reveals a scarcity in large-scale, freely accessible datasets that meet comprehensive research needs.

The most notable, publicly available resources in Ukrainian include:

1. BRUK (Starko and Rysin, 2023), a corpus of 450,000 words, whose genre distribution mirrors that of the original Brown corpus[3], covering fiction, religious texts, press, legal documents, etc.;

2. UA-GEC (Syvokon et al., 2023), a corpus of 500,000 words, which contains texts with errors and their corrections from a wide variety of writing domains, from text chats and essays to formal writing;

3. UberText 2.0 (Chaplynskyi, 2023), which consists of 8.59 million texts of news, fiction, social media posts, Wikipedia, and court decisions;

[1] https://github.com/Stereotypes-in-LLMs/recruitment-dataset

[2] https://djinni.co/

[3] http://korpus.uib.no/icame/manuals/brown/

4. Malyuk[4], a corpus of 38.94 million texts, which is a compilation of UberText 2.0, Oscar[5] (derived from Common Crawl), and Ukrainian News[6];

5. UD Ukrainian[7], a gold standard Universal Dependencies corpus for Ukrainian, which comprises 7,000 sentences of fiction, news, opinion articles, Wikipedia, legal documents, letters, posts, and comments.

Despite the genre diversity present in the publicly available corpora for Ukrainian, none of them include texts from the recruitment domain.

In our search for open-source job-related datasets, we identified relevant corpora for the English language, but they focus on either job descriptions[8] or candidate CVs[9], without offering a unified set that would cater to both aspects. This disjointed approach inhibits the capability to perform semantic matching, thereby constraining the development of automated job recommender and AI-assisted hiring systems.

The corporate landscape of open-source datasets is similarly fragmented: platforms like Indeed[10] provide separate datasets for CVs[11] and job descriptions[12]. Structural and temporal differences in these datasets challenge the development of NLP models for effective job-candidate matching. This situation emphasizes the need for more collaborative efforts between academia and industry to foster the creation of open, integrated datasets.

3. Dataset Description

In this section, we'll delve into the Djinni Recruitment Dataset, detailing its structure and processing and offering key insights into notable features.

3.1. Data Source

The data in the corpus originates from Djinni, Ukraine's leading tech job marketplace, boasting over 50,000 monthly users. Djinni generously provided open-source access to two significant data groups: anonymous candidate information and job descriptions, primarily from the IT sector in Ukraine. This wealth of data serves as a valuable resource for understanding trends and patterns in the Ukrainian tech job market. In the pursuit of accurate analysis, we conducted additional preprocessing of this data, a topic we explore further in the next section.

3.2. Data Processing

The dataset underwent several critical preprocessing steps, including language filtering, the filtering of duplicates and outliers, language-based split, and the removal of personally identifiable information.

3.2.1. Data Filtering

We used the langdetect[13] model from the transformers library[14] to detect and select data samples exclusively in English and Ukrainian languages, ensuring the dataset's relevance to the primary language groups within the Ukrainian IT sector.

To improve the dataset's diversity and balance, we undertook a deduplication effort, focusing on removing both exact duplicates and highly similar samples. We employed embedding models to identify similar CVs and job descriptions, selecting models based on their quality for each language at the time of filtering. Specifically, for English texts, we used the bge-base-en-v1.5 model[15] with an empirically determined cosine similarity threshold of 0.9. For Ukrainian texts, we chose the multilingual-e5-large model[16] with the threshold of 0.95. This approach ensured that the dataset comprised only high-quality, unique entries.

Moreover, we implemented an outlier removal step, filtering out entries below the 5th percentile in text length to exclude extremely short texts. This refinement enhances the dataset's relevance.

We monitored the impact of the filtering on the size of the dataset at each filtering stage. Table 1 shows that candidate CVs experienced a modest reduction of 20%, whereas job descriptions saw

[4] https://huggingface.co/datasets/lang-uk/malyuk

[5] https://huggingface.co/datasets/oscar

[6] https://huggingface.co/datasets/zeusfsx/ukrainian-news

[7] https://github.com/UniversalDependencies/UD_Ukrainian-IU/tree/master

[8] https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset, https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/data

[9] https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

[10] https://www.indeed.com

[11] https://datastock.shop/download-indeed-job-resume-dataset/

[12] https://data.world/promptcloud/indeed-job-posting-dataset

[13] https://huggingface.co/ERCDiDip/langdetect

[14] https://huggingface.co/docs/transformers/en/index

[15] https://huggingface.co/BAAI/bge-base-en-v1.5

[16] https://huggingface.co/intfloat/multilingual-e5-large

	CVs	Jobs
Raw samples	294,678	443,458
After basic filtering	241,561	358,491
After similarity filtering	234,480	169,358

Table 1: The number of samples in the dataset before and after filtering. Basic filtering includes language filtering and the removal of outliers and identical duplicates. Similarity filtering covers the removal of near-identical samples.

a more substantial decrease of 60%. The cause of this contrast lies in the highly repetitive nature of job descriptions posted by the same companies in different periods, which we verified via a closer data analysis.

3.2.2. Language-Based Split

We split the dataset into two based on the detected language, forming separate divisions for English and Ukrainian sections within both job descriptions and CVs. This strategic division enables more nuanced analysis and application of NLP techniques tailored to language specifics, significantly enhancing the relevance of insights derived from the dataset for bilingual environments. This step also revealed a serious imbalance of language representation: Ukrainian-language CVs constitute only 10% of all CVs, and Ukrainian-language job postings constitute 16% of all job postings. The exact numbers can be found in Table 2.

	CVs	Jobs
English	210,250	141,897
Ukrainian	24,230	27,461

Table 2: The number of CVs and job descriptions in the English and Ukrainian segments of the dataset post language-based splitting.

3.2.3. Removal of Personally Identifiable Information

Djinni has a strict policy requiring registration through anonymized profiles only and enforces measures to prevent the posting of personally identifiable information (PII). This approach to anonymity ensures the protection of sensitive personal data and reduces bias during resume screening by potential employers.

To verify the anonymity and confidentiality of the dataset, we developed a script[17] utilizing regex implementation tailored for both English and Ukrainian languages. The script is based on patterns and keywords in both languages, covering phone numbers, email addresses, physical addresses, social media links, taxpayer identification numbers, and other unique identifiers. This step was pivotal in detecting remnants of PII within CVs.

The identified CVs with PII were meticulously removed from the dataset to uphold the highest standards of privacy and data protection. Less than 0.2% of the CVs contained PII data.

For further details on the attributes of the CV and job description datasets, see Appendix A: Feature Explanation.

4. Protected Attributes in the Dataset

Our research further focused on identifying protected attributes within the anonymized CVs to determine the true level of anonymity in the provided data, as well as to pinpoint potential sources of bias in recruitment practices. Following the Principles of Preventing and Combating Discrimination[18] in Ukraine, we identified core protected attributes for our study: gender, age, marital status, military status, religion, and person name.

Our analysis primarily focused on identifying explicit mentions of protected attributes in CVs across both English and Ukrainian languages. We developed a script that uses regular expressions and dictionaries to detect terms and patterns related to specific protected attributes[19]. To detect person names, we used the VESUM[20] dictionary, which contains more than 5 thousand names in Ukrainian, and translitua[21] to transliterate Ukrainian names and enable search in the English segment. We manually crafted parallel dictionaries in both Ukrainian and English for other protected attributes: 22 gender groups, ages from 16 to 65 years, 5 marital statuses, 5 military statuses, and 9 religious groups. The script can be used to improve data anonymity and increase fairness in automated hiring processes.

4.1. Experimental Findings

The quantitative insights into the explicit representation of protected attributes within the dataset, categorized by language, are presented in Table 3.

[17] https://github.com/Stereotypes-in-LLMs/recruitment-dataset/blob/main/notebooks/EDA/PII_CV_analyses.ipynb

[18] https://zakon.rada.gov.ua/laws/show/5207-17

[19] https://github.com/Stereotypes-in-LLMs/recruitment-dataset/blob/main/notebooks/EDA/EDA_candidates.ipynb

[20] https://github.com/brown-uk/dict_uk

[21] https://pypi.org/project/translitua/

Protected Group	Ukr CVs (%)	Eng CVs (%)
Age	**0.21**	0.15
Gender	**0.66**	0.05
Marital Status	**0.07**	0.02
Military Status	**0.42**	0.26
Name	3.75	**3.85**
Religion	0.02	**0.2**

Table 3: The fractions of CVs that contain explicit mentions of protected attributes.

This analysis reveals significant differences between Ukrainian and English CVs. Particularly, explicit mentions of gender are substantially more frequent in Ukrainian CVs, while mentions of religion are much more common in English CVs. The results show that beyond PII, certain characteristics may introduce bias, necessitating their anonymization for the further use of the dataset.

4.2. Gender-Marked Verbs in Ukrainian CVs

Unlike English, Ukrainian is a synthetic language, whose verbs are inflected for the grammatical gender when used in the past tense. This means that an anonymous CV that uses gender-marked verbs may reveal the gender of the author.

To analyze the impact of this linguistic phenomenon, we developed a script[22], which uses the pymorphy3[23] and stanza[24] Python libraries to analyze texts. In each Ukrainian CV, we then identified gender-marked verbs, which related to the subject "I" or had no subject, and checked which grammatical gender prevailed, subsequently classifying those CVs as revealing the author's gender.

The proposed metric allowed us to detect 16.55% of Ukrainian CVs that may have been written by candidates who identify as female and 30.50% by candidates who identify as male. This analysis highlights the nuanced ways gender perspectives may be integrated into job-related documents and emphasizes the need for more elaborate strategies for detecting protected attributes. We leave this for future work.

5. Intended Use

The Djinni Recruitment Dataset can be leveraged for the purposes outlined below:

1. for the development of recommender systems and advanced semantic search;
2. as potential training data for both English and Ukrainian domain-specific LLMs, based on GPT-3 (Brown et al., 2020), Llama 2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Palm 2 (Anil et al., 2023), etc., enriching their understanding and generating capabilities within specialized recruitment contexts;
3. as a benchmark or training set to promote fairness in AI-assisted hiring, addressing bias and ensuring equitable selection processes;
4. for automated resume and job description creation;
5. for market analysis and evaluation of the tech sector's dynamics in Ukraine;
6. for topic discovery and trend analysis within the tech industry through modeling and classification;
7. for automated identification of company domains, assisting in strategic market planning.

6. Challenges and Limitations

We acknowledge the following challenges and limitations of the Djinni Recruitment Dataset:

1. **Limited languages:** The dataset is available in only two languages—Ukrainian and English.
2. **Unlabelled data:** The lack of labeled data makes it challenging to determine who was hired and to conduct specific analyses related to successful job placements.
3. **Lack of CV publication date:** The dataset does not include any information on when the CVs were published.
4. **Noisy user-generated data:** The dataset includes user-generated content, introducing noise and variability that may impact the accuracy of certain analyses.
5. **Focus on the tech domain:** The dataset is primarily centered around the tech domain, limiting its applicability to other industries or sectors.
6. **Ukrainian market only:** The dataset exclusively represents the Ukrainian market, which may restrict broader generalizations or comparisons with job markets in other regions.

Understanding these challenges is crucial for the appropriate interpretation and utilization of the dataset in a way that aligns with its inherent limitations.

[22] https://github.com/Stereotypes-in-LLMs/recruitment-dataset/blob/main/notebooks/EDA/EDA_candidates.ipynb

[23] https://pypi.org/project/pymorphy3/

[24] https://stanfordnlp.github.io/stanza/

7. Conclusion

In this paper, we introduced the Djinni Recruitment Dataset, a pioneering resource in NLP and recruitment data analysis, with a focus on the Ukrainian IT sector, which contains data in the Ukrainian and English languages. The dataset is released under the MIT license, which allows for academic and commercial use.

This dataset's focus on recruitment is key for creating NLP tools for job matching, market analysis, bias identification, and fostering Responsible AI in hiring. Its bilingual content represents the tech sector of Ukraine, largely influenced by the global IT job market.

One of the most significant contributions of the Djinni Recruitment Dataset is that it sets a precedent for other businesses to consider the value of making their data openly available for research purposes.

Future research may expand the dataset's languages and industries. There's potential for creating targeted NLP tools to improve recommendation systems and algorithms for bias detection and mitigation in the recruitment domain.

8. Ethical Considerations

The Djinni Recruitment Dataset adheres to the conditions of fair use. The contributors of data have the privilege to ask for their information to be deleted by contacting the authors of this paper.

The Djinni dataset upholds standards of data anonymization and privacy protection. These measures are implemented to prevent any potential harm to the authors of the data. By prioritizing anonymity, we strive to safeguard the privacy of those who have contributed to this valuable resource.

The dataset is published with the description of intended use, which underscores our commitment to responsible data stewardship.

We used ChatGPT and Grammarly to assist with paraphrasing while writing this paper.

9. Acknowledgments

We express our gratitude to Djinni for sharing the data. Djinni's dedication to advancing knowledge and fostering progress in the AI-driven recruitment field has enabled the creation of this dataset.

10. References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeffrey Dastin. 2018. Insight - amazon scraps secret ai recruiting tool that showed bias against women.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. pages 96–102.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, and Guillem Cucurull. 2023. Llama 2: Open foundation and fine-tuned chat models.

A. Feature Explanation

A.1. Job Descriptions

Both English and Ukrainian parts of the dataset contain attributes related to job descriptions, including position titles, job descriptions, company names, experience requirements, keywords, English proficiency levels, publication dates, language of job descriptions, and unique identifiers.

Features:

- **id:** 169,358 unique synthetic identifiers for each job description.
- **Position:** 82,423 unique manually written position titles.
- **Long Description:** 169,358 unique manually written job descriptions.
- **Company Name:** 12,897 unique company names.
- **Exp Years:** 5 unique values for experience years required: ’2y’, ’3y’, ’no_exp’, ’5y’, ’1y’.
- **Primary Keyword:** 46 unique job profile types.
- **English Level:** 6 unique English proficiency levels: ’intermediate’, ’pre’, ’upper’, ’basic’, ’fluent’, NaN.
- **Published:** publication dates (only month and year).
- **Long Description_lang:** 2 unique languages in which job descriptions can be written: ’uk’ (Ukrainian), ’en’ (English).

A.2. CVs

Both English and Ukrainian parts of the dataset contain attributes related to candidate CVs, including position titles, candidate information, candidate highlights, job search preferences, job profile types, English proficiency levels, experience years, concatenated CV text, language of CVs, and unique identifiers.

Features:

- **id:** 234,480 unique synthetic identifiers for each candidate CV.
- **Position:** 58,341 unique manually written position titles.
- **Moreinfo:** 234,365 unique manually written candidate information entries.
- **Looking For:** 109,524 unique manually written job search preferences.
- **Highlights:** 117,700 unique manually written candidate highlights.
- **Primary Keyword:** 42 unique job profile types.
- **English Level:** 7 unique English proficiency levels: ’intermediate’, ’pre’, ’upper’, ’basic’, ’no_english’, ’fluent’, NaN.
- **Experience Years:** 15 unique values representing candidate experience in years.
- **CV:** 234,480 unique concatenated CV texts (Highlights + Moreinfo + Looking For).
- **CV_lang:** 2 unique languages in which CVs can be written: ’uk’ (Ukrainian), ’en’ (English).