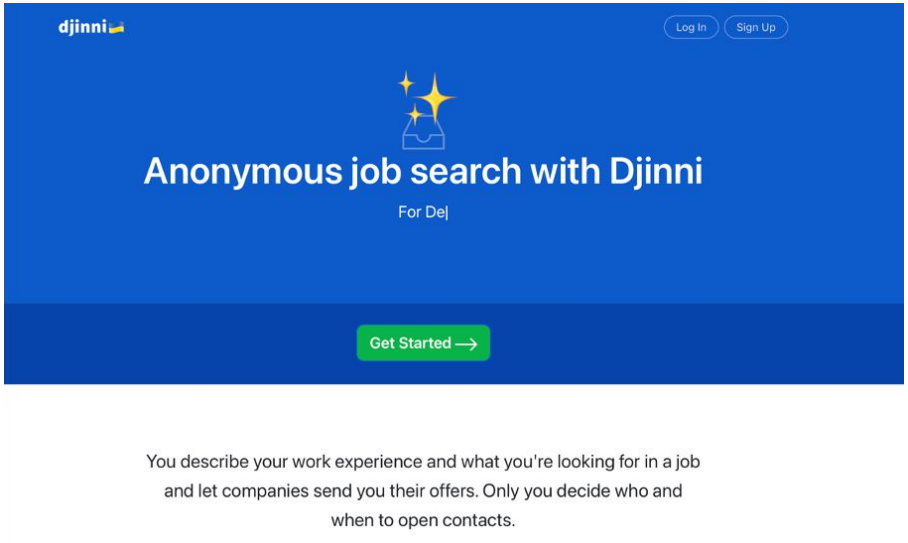
Introducing the Djinni Recruitment Dataset: A Corpus of Anonymized CVs and Job Postings

Authors: Nazarii Drushchak, Mariana Romanyshyn

Acknowledgements

Special thanks to the **Djinni team** for providing the data.



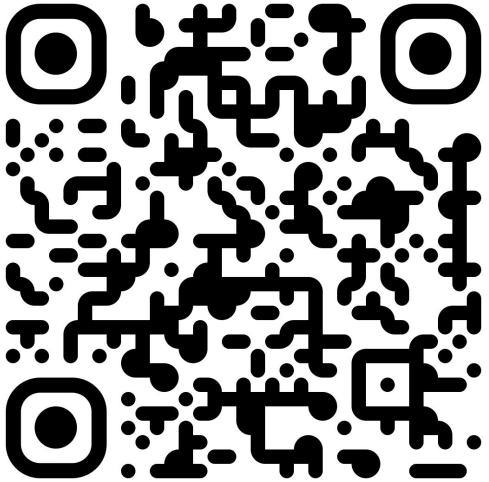
Agenda

- Introduction and Motivation
- Dataset Filtering and Statistics
- Data Analysis
- Challenges and Limitations
- Intended Use

Introduction and Motivation

Djinni Recruitment Dataset

- Large-scale open-source corpus of job descriptions and candidate profiles;
- Over 150k jobs and 230k candidates in English and Ukrainian;
- Facilitates NLP advancements in recruitment;
- MIT license.



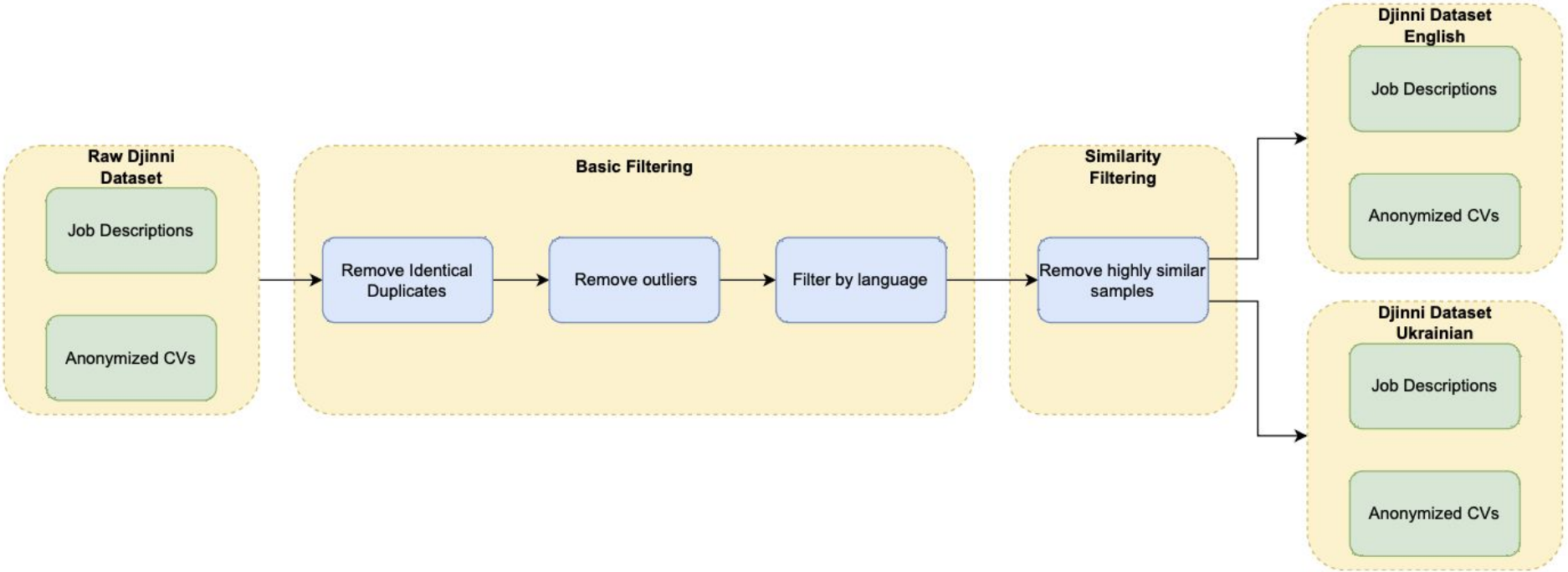
Github Repository

Why did we do this research?

- Limited availability of diverse, publicly accessible datasets in the recruitment domain, particularly for low-resource languages like Ukrainian;
- Addressing the scarcity of integrated datasets covering both job descriptions and candidate profiles;
- Contributing to the discourse on Responsible AI practices and fairness in hiring systems.

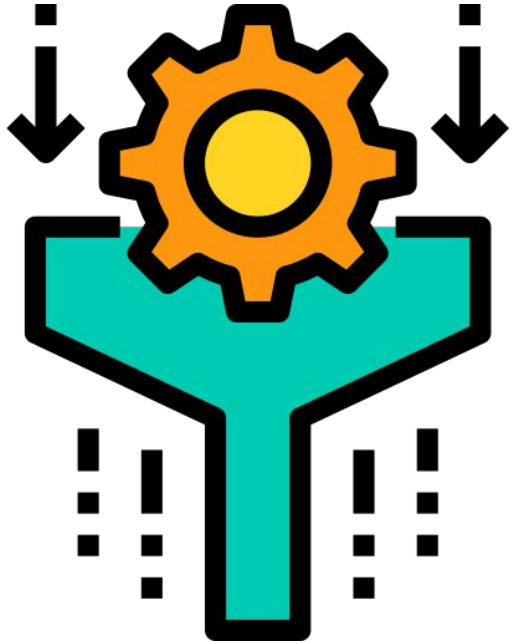
Dataset Filtering and Statistics

Dataset Processing/Filtering Flow



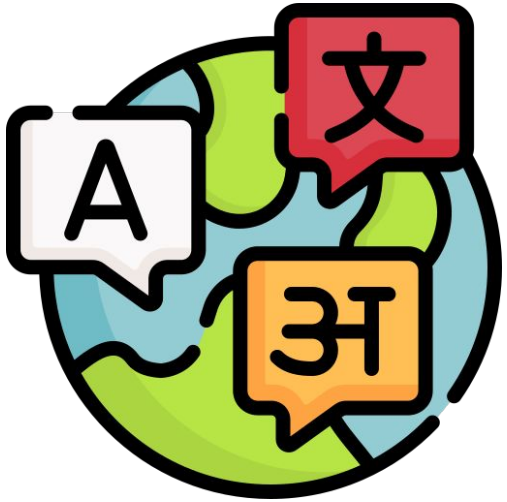
The Number of Samples Before and After Filtering

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



Language-Based Split

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



Data Analysis

What is protected group?

Protected Group refers to individuals or categories safeguarded by law against discrimination based on specific characteristics like gender, age, marital status, military status, religion, and person name, etc.

Protected Attributes in Dataset

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**

Funny names we filtered out

- Main
- Hi
- Feb
- Pet
- Apollo
- etc.

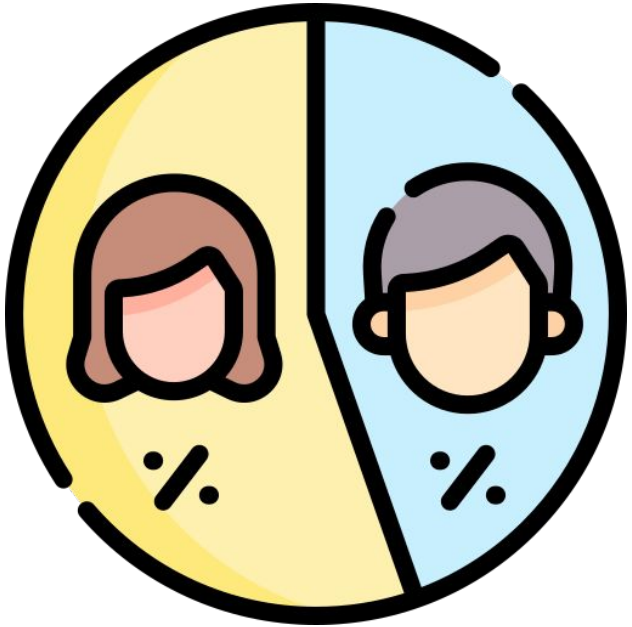
Funny names we filtered out

	Ukr version
Main	Майн
Hi	Гі
Feb	Феб
Pet	Пет
Apollo	Аполло
etc.	

VESUM Dictionary

Gender-Marked Verbs in Ukrainian CVs

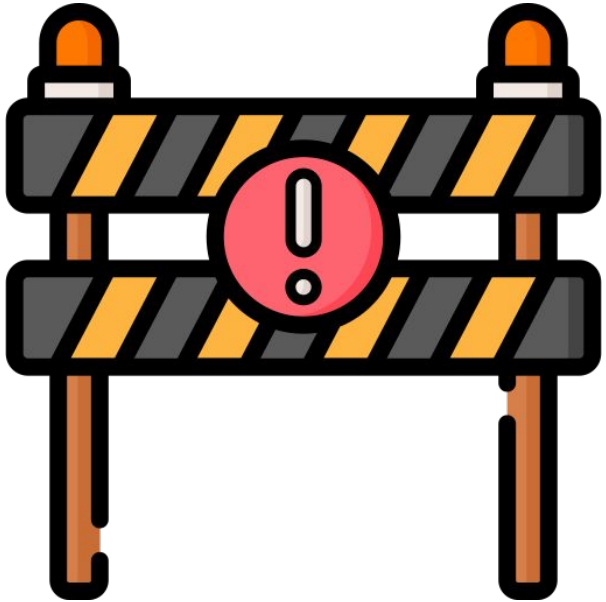
- **16.55%** of Ukrainian CVs that may have been written by candidates who identify as **female**;
- **30.50%** by candidates who identify as **male**.



Challenges and Limitations

Challenges and Limitations

- Limited languages;
- Lack of CV publication date;
- Noisy user-generated data;
- Focus on the tech domain;
- Ukrainian market only.



Intended Use

Intended Use

- for the development of recommender systems and advanced semantic search;
- as potential training data for both English and Ukrainian domain-specific LLMs;
- as a benchmark or training set to promote fairness in AI-assisted hiring, addressing bias and ensuring equitable selection processes;
- *Check more in the paper.*

Q&A