A machine-learning based model to identify PhD-level skills in job ads

Australian National University

*Contact:
li'an.chen@anu.edu.au

Li'An Chen

Hanna Suominen

Inger Mewburn

Acknowledgement

We are grateful for the support from Burning Glass Technologies Inc, PostAc, and ANU CV Discovery Translation Fund 2.0, nonymous paper reviewers, Dr. Will, Dr. Lindsay Hogan, Chenchen, the anonymous annotator's and Australian Government Research Training Program International Scholarship.

Introduction

Nowadays, almost 60% PhD graduates end up working in industry. Although many call for measures to more closely align PhD curriculum with the needs of industry, current PhD programs still focus exclusively on research skills. Reports have shown industry employers were not satisfied with PhD graduates' employability skills. Whilst surveys and interviews have been used to extract employers' expectation of PhD graduates, this study explores a new approach of using ML-NLP techniques to directly extract skill requirements from research skill intensive job ads.

Objectives

- Outline a robustly developed coders' workshop valuable for future ML-NLP attempts to inform higher education policy making
- Provide curriculum design tools for specifying the degree to which a skill is required of PhD graduates
- Testify the hypothesis that contextual difference exist across industry domains

Methods

Human analytic framework: Move-step analysis

- Move: Coarse-grained items
- Steps: Fine-grained items

Coders' workshop

- Multiple coders (2)
- Iterative process of continuous improvement
- Mechanism to resolve interpersonal conflicts
- Inter-coder reliability - Cohen's Kappa = 0.76

Data

400 research intensive job ads in Health & Computing (147,089 words)

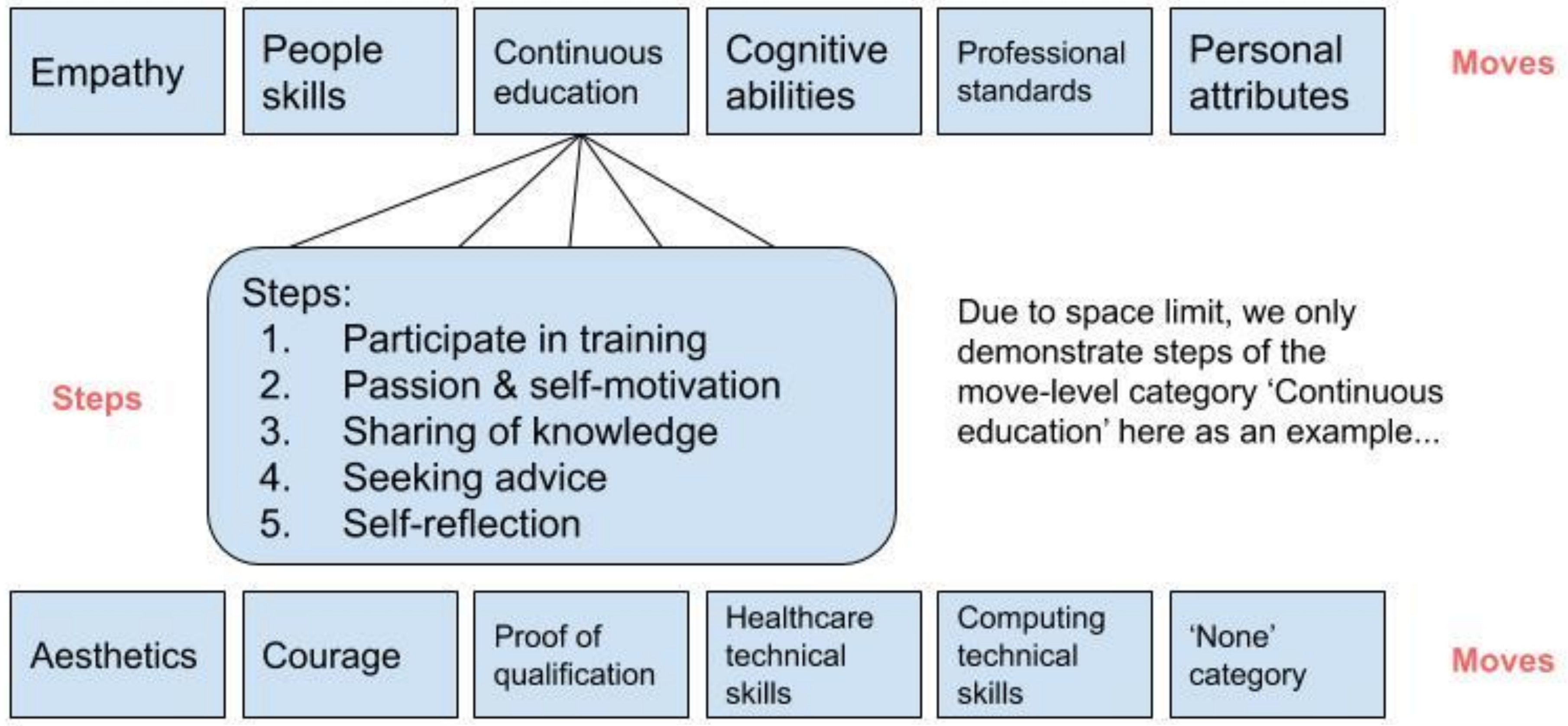
ML algorithm: Linear SVM

- Suitable for multiclass classification
- Computationally cost-effective
- Give good prediction outcomes for similar tasks

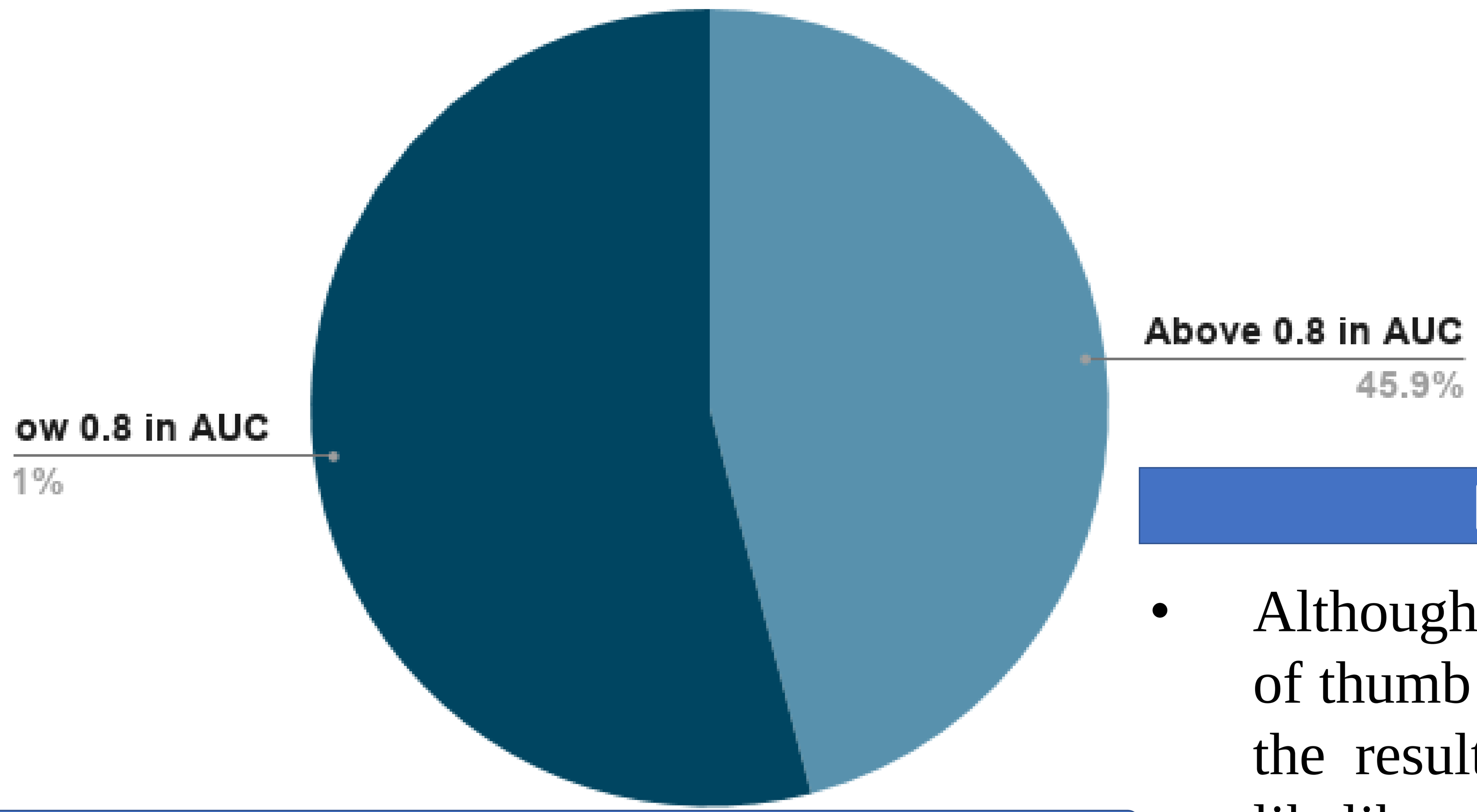
Performance evaluation: Area Under the Curve (AUC)

- Avoid situation where accuracy score is high, but evaluation is biased by class imbalance

Results



Ratio of step-level categories with varied ML performance



- 12 categories identified at move-level
- 61 step-level categories associated with move-level skills
- 28 step-level categories reached AUC score above 0.8 on training, validation, test sets.
- Significant difference between the two industry domains in 22 of these 27 step-level categories (excluding 'None') that reached good AUC score

Limitation

- Before the model can be used to automatically identify all the skill categories, the overfitting and underfitting problems need to be solved.
- we only experimented with limited number of data in healthcare and computing domains. Therefore, extra effort to manually label more data or cross-validation approach can be conducted in the future to improve the experiment process.

Discussion & Conclusion

- Although not all the skill categories reached the rule of thumb gold standard of 0.8 in AUC performance, the results from the experiment so far indicate a likelihood that the manual efforts can be replicated by the machine when further optimization is conducted.
- Through Chi-square test, we pinpointed difference in the number of skills required in computing and healthcare industries. These difference challenges the view that doctoral graduates' identity is monolithic.
- Our human coder's workshop experience can be useful for scholars who also intend to conduct ML driven analysis of textual data for enabling better decision making in higher education.