Enhanced Resume Screening for Smart Hiring: Integrating Multi-Class Classification, Fine-Grained Skill Extraction, and End-to-End Pipeline Integration

Shaurya Jain
Rutgers University
Email: sj1029@rutgers.edu

Pavan Madamsetty
Rutgers University
Email: pm950@rutgers.edu

Reo Correia
Rutgers University
Email: rc1460@rutgers.edu

Abstract

We propose an enhanced resume screening system that integrates multi-class classification, fine-grained skill extraction, and an end-to-end pipeline for interactive smart hiring. Our system extends prior S-BERT–based similarity ranking methods by classifying resumes into 22 predefined job categories and extracting candidate skills via a robust Named Entity Recognition (NER) pipeline. Preliminary experiments on a dataset of over 2400 resumes indicate significant improvements in both candidate categorization and skill identification, promising a more efficient and accurate resume screening process for modern HR applications.

1 Introduction and Problem Statement

Automated resume screening is increasingly vital due to the large volume of applications received by organizations. Existing methods, such as S-BERT–based resume ranking (Reimers and Gurevych, 2019), primarily focus on measuring the similarity between resumes and job descriptions. However, these approaches lack explicit multi-class categorization and detailed, fine-grained skill extraction.

Our project addresses these shortcomings by developing an integrated system that:

- Classifies resumes into 22 predefined job categories (e.g., HR, IT, Finance, etc.) using a fine-tuned transformer-based classifier.
- Extracts and normalizes candidate skills through an NER pipeline built with tools such as spaCy (AI, 2020) and Hugging Face Transformers (Devlin et al., 2019).
- Combines these components into an end-to-end system with a user-friendly interface, enabling HR managers to efficiently review and query candidate profiles.

2 Data

Our primary dataset consists of over 2400 resumes available in PDF and text formats. Accompanying the resumes is a CSV file that contains:

- **ID**: Unique identifier for each resume.
- **Resume_str**: Text extracted from resumes.
- **Resume_html**: HTML representations obtained during web scraping.
- **Category**: The job category applied for (across 22 predefined categories).

In addition, we plan to augment the dataset with corresponding job descriptions for each category. Preprocessing will involve text extraction, noise reduction, and manual annotation for skill entities to ensure data quality and consistency.

3 Methodology and Initial Experiments

Our approach is divided into three main components:

3.1 Multi-Class Resume Classification

We will fine-tune a pre-trained transformer model (e.g., DistilBERT) on the labeled resume dataset. The goal is to accurately classify resumes into 22 job categories using standard metrics such as accuracy, precision, recall, and F1-score. Our initial experiment involves training a baseline model with an 80/20 train-test split to evaluate its performance.

3.2 Skill Extraction via NER

We will build an NER pipeline using spaCy and Hugging Face Transformers to extract candidate skills from resume text. A custom skills dictionary and normalization scheme will be developed to ensure that variant forms of skills (e.g., "C++", "C plus plus") are unified. Initial experiments will compare the automated skill extraction against a manually annotated ground truth.

3.3 End-to-End Pipeline and User Interface

Finally, we will integrate the classification and skill extraction modules into a single pipeline using orchestration frameworks (such as LangChain). A simple web interface will be developed with Streamlit to allow HR managers to upload resumes, view predicted job categories, and see the extracted skills in real-time.

4 Timeline and Expected Contributions

Timeline:

- **Week 1:** Data preprocessing and baseline classifier training.
- **Week 2:** Development of the NER-based skill extraction pipeline.
- **Week 3:** Integration of modules and prototype development of the web interface.
- **Week 4:** Extensive evaluation, documentation, and preparation of the final presentation.

Expected Contributions:

- A transformer-based classifier that improves resume categorization into 22 job categories.
- A robust NER pipeline for extracting and normalizing candidate skills.
- An integrated and interactive resume screening system that significantly enhances the hiring process.

5 Conclusion

Our proposed system builds upon prior S-BERT–based methods by incorporating explicit multi-class classification and detailed skill extraction. By integrating these modules into an end-to-end solution, we expect to enhance both the efficiency and accuracy of automated resume screening, offering valuable advancements for modern HR practices.

Limitations

This proposal outlines a system built on available NLP techniques and open-source tools. Limitations may include dependency on the quality of resume text extraction from PDFs and the generalizability of the classifier to unseen categories or domains. Further work is required to mitigate potential biases in automated screening.

Acknowledgments

We thank our advisors and collaborators for their valuable feedback and support during the development of this project proposal.

References

Explosion AI. 2020. spacy: Industrial-strength natural language processing in python. https://spacy.io/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.