Progress report: *Resume Classification and Ranking*

Mian Afrasiyyab Farakh
BSCS – 10 - C
CMS: 335942
SEECS , NUST
Islamabad, Pakistan
mfarakh.bscs20seecs@seecs.edu.pk

Kamlesh Kumar
BSCS – 10 - C
CMS: 348395
SEECS , NUST
Islamabad, Pakistan
kkumar.bscs20seecs@seecs.edu.pk

Usama Qureshi
BSCS – 10 - C
CMS: 338954
SEECS , NUST
Islamabad, Pakistan
uqureshi.bscs20seecs@seecs.edu.pk

Abstract

This progress report presents an ongoing project focused on developing a deep learning (DL) model for automating the resume review process. The project aims to address the challenges faced by human resource departments in screening and shortlisting candidates for job openings. At this stage of project, we have successfully implemented classification model using transformers, enabling the classification of resumes based on education, work experience, skills, and achievements. The next stage involves utilizing the ML model for resume ranking within classes, considering factors such as skills, education and overall quality. The application of the ML model for resume classification and ranking has broad potential across industries, streamlining talent acquisition and enhancing recruitment efficiency. This progress report highlights the completion of the classification task and outlines the next steps towards achieving automated resume ranking.

KEYWORDS- *MACHINE LEARNING MODEL, NLP, RESUME PARSING, REGEX, NAMED ENTITY RECOGNITION (NER), PYRESPARSER, K-NN, SVM, NAÏVE BAYES, TRANSFORMER MODELS, COSINE SIMILARITY, WORD CLOUD, MODEL OPTIMIZATION.*

I. INTRODUCTION

Motivation: The traditional process of reviewing resumes for job openings is highly time-consuming and prone to biases. Human resource departments often receive a large number of resumes for each job opening, making it challenging to manually screen and shortlist the most suitable candidates. Additionally, the lack of a standard structure and format for resumes further complicates the process, requiring domain knowledge to assess the relevance and applicability of each profile. These inefficiencies can lead to missed opportunities for both job seekers and employers, and there is a clear need for an automated solution to streamline the resume review process.

Our project aims to address these challenges by developing a machine learning model for resume classification and ranking. By leveraging the power of Machine learning, natural language processing, and transformers, we intend to create a system that can automate the initial screening of resumes, effectively shortlisting the most relevant candidates for a specific job opening.

This will not only save valuable time for human resource professionals but also improve the overall efficiency and effectiveness of the recruitment process.

Possible Applications: The application of our ML model for resume classification and ranking is not limited to any specific industry or domain. It can be utilized across various sectors, including IT, finance, healthcare, manufacturing, and many others. Regardless of the field, companies and organizations constantly face the challenge of selecting the most suitable candidates from a large pool of applicants. Our system can significantly enhance the recruitment process by automating the initial resume screening, enabling HR departments to focus their attention on more strategic and value-added tasks.

Furthermore, our model can be tailored to specific job roles and requirements, allowing companies to efficiently identify candidates with the desired skills, education, and experience. Whether it's identifying software engineers with expertise in a particular programming language or selecting project managers with a proven track record, our ML model can effectively analyze and classify resumes based on specific criteria.

II. BACKGROUND AND DOMAIN KNOWLEDGE

The process of reviewing and ranking resumes is a key task within the domain of Human Resource Management (HRM) and Recruitment. It is a multifaceted problem that encompasses natural language processing (NLP), data analysis, machine learning (ML), and domain-specific knowledge in recruitment. The project "Resume Classification and Ranking" intersects with these disciplines to create an efficient and automated solution for resume screening.

Historically, the task of resume screening has been performed manually, resulting in a time-consuming and often biased process. However, the advent of machine

learning and NLP technologies has opened new avenues for automating and improving this process. Previous studies have demonstrated the potential of these technologies in the context of resume screening and candidate shortlisting [1-2].

Our project falls under the broader field of Information Extraction (IE) in NLP, where the goal is to extract structured information from unstructured text data. In the specific context of our project, the unstructured text data is the resumes, and the structured information includes the qualifications, skills, experience, and other attributes of the job applicants.

The classification and ranking of resumes are essentially tasks of document classification and information retrieval. We leverage transformer-based models, a breakthrough in the field of machine learning and NLP, known for their effectiveness in understanding the context of language. Transformers, through their attention mechanisms, can understand the context of words in a sentence, making them highly efficient for tasks such as text classification, sentiment analysis, and named entity recognition, among others.

For the ranking aspect, we employ similarity measures such as cosine similarity, a popular choice for comparing document similarity in high-dimensional spaces. By transforming resumes and job descriptions into vector representations (embedding), we can compute the cosine similarity to quantify how closely a candidate's profile matches the job requirements.

III. Literature Review and Project Relevance

The field of automated resume classification and ranking is an active area of research, given its significance in improving the efficiency and fairness of the recruitment process. Several studies have highlighted the potential of machine learning and natural language processing technologies in this domain.

In the work of Zhang (2020), a deep learning approach was used to extract information from resumes [1]. Dahiya and Singh (2017) employed natural language processing and clustering for resume ranking [2]. Other researchers have used recommender systems for candidate shortlisting [3]. These studies validate the relevance of our project to current research trends.

However, our project distinguishes itself from existing solutions in several ways. Firstly, we leverage transformer-based models for resume classification. Transformers, introduced by Vaswani et al. (2017), are relatively new and represent a significant shift in the field of NLP, particularly due to their self-attention mechanism that helps in understanding the context of language [4]. Despite their promising potential, transformer models are under-explored in the context of resume classification and ranking, and our project aims to fill this gap.

Secondly, most existing solutions focus either on resume classification or ranking. Our project, on the other hand, aims to address both aspects, creating a more holistic solution. We not only classify resumes based on their relevance to a job opening but also rank them based on the degree of match with job requirements.

Therefore, our project aligns with the current research trends while also offering novel approaches and improvements over existing solutions.

IV. Preliminary Experiments and Results

Our preliminary experiments involved parsing resumes to extract relevant fields, classifying resumes into different professional categories, and ranking the resumes within a single category.

Resume Parsing: In the initial stage, we attempted to extract relevant fields such as education, professional experience, and skills from resumes. We experimented with various approaches, including regular expressions (regex), Named Entity Recognition (NER) models, and third-party resume parsers.

Our initial efforts with regex, although straightforward, proved insufficient due to its limitations in handling complex and varied resume structures and terminologies. Subsequently, we experimented with NER using the "en-core-web-sm" model from the Spacy library. While NER proved more effective than regex, it still faced challenges with varying resume formats and synonymous terminologies[5]. Our efforts with third-party resume parsers, specifically python "pyresparser", showed promise and we are further refining our parsing techniques to increase accuracy and consistency .

Resume Classification: Having parsed the resumes, we proceeded with the classification task. We experimented with traditional machine learning approaches including K-Nearest Neighbors (K-NN), Support Vector Machines (SVM), and Naïve Bayes. However, these methods did not yield satisfactory results, potentially due to the high-dimensional and sparse nature of text data, as well as the potential non-linearity of the classification boundaries.

Motivated by the recent success of transformers in NLP tasks, we implemented a transformer-based model for the classification task. The results were significantly better than traditional machine learning approaches. After just three epochs, we achieved an accuracy of 73%, a promising result that motivated us to further explore and optimize this approach.

Resume Ranking: The ranking of resumes within a single category is still under experimentation, and we plan to leverage our classification results and cosine similarity measures for this task.

V. Problems and Potential Solutions

Problem 1: Resume Parsing: The first challenge we encountered was extracting relevant fields, such as education, professional experience, and skills, from resumes. The different resume parsing approaches that we tested—regex, Named Entity Recognition (NER), and third-party parsers—all had their limitations. The regex approach was too simplistic, while NER struggled with different resume formats and terminologies. Third-party parsers, despite being larger in size and diverse in language implementation, showed potential. We explored various parsers and finally decided to utilize "pyresparser," which delivered promising results. Future work will involve further refining and optimizing our resume parsing techniques [5].

Problem 2: Resume Classification: Another challenge was categorizing resumes into different professional fields. We reviewed several studies that employed various approaches, from machine learning models to deep learning techniques. However, traditional machine learning methods like K-NN, SVM, and multinomial Naïve Bayes resulted in low accuracy. Inspired by the success of deep learning in NLP tasks, we decided to adopt a transformer-based model for classification [6]. This approach has yielded significantly better results in our preliminary experiments, and we plan to continue optimizing it.

Problem 3: Resume Ranking: The final challenge we foresee is the ranking of resumes within a single professional category. We have not yet tackled this aspect of the project but have studied potential approaches. Our current plan is to construct a word cloud for each profession category. The cosine similarity between the skill set extracted from a resume and the word cloud for the corresponding profession will serve as a measure for ranking. Additionally, we plan to assign weights to specific educational qualifications, like bachelor's and master's degrees, which will contribute to the final ranking. Future work will involve implementing and testing this approach [7].

We have identified three main challenges in our project: resume parsing, classification, and ranking. We have made progress in addressing the first two challenges and have a potential approach for the third. We believe that by further refining our methods and integrating modern NLP techniques, we can effectively overcome these challenges.

V.I. Objectives Met and Pending

In our project "Resume Classification and Ranking," we have identified several objectives to ensure the successful development and application of our machine learning model. Here, we discuss the objectives we have met and those we are yet to achieve.

A. Objectives Met:

1. **Resume Parsing**: We have explored several techniques for parsing resumes and extracting relevant information, including regex, Named Entity Recognition (NER), and third-party parsers. Although each technique had its limitations, we have finally selected "pyresparser" for its promising results and accuracy.

2. **Resume Classification**: We have investigated several machine learning and deep learning approaches for classifying resumes into different professional categories. Despite the low accuracy of traditional machine learning methods, we have successfully implemented a transformer-based model that shows significant promise, achieving an accuracy of 73% in our preliminary experiments.

3. **Literature Review**: We have extensively reviewed the literature on resume classification and ranking. Our review confirms that our project aligns with current research trends, while also contributing novel approaches to the field.

B. Objectives Yet to be Met:

1. **Resume Ranking**: We have yet to work on the ranking of resumes within a single professional category. Our preliminary plan involves creating a word cloud for each profession category and using cosine similarity to rank resumes. Additionally, we aim to assign weights to certain educational qualifications. This aspect of the project remains to be implemented and tested.

2. **Model Optimization**: While our transformer-based model for resume classification has shown promise, we believe there is room for further optimization. We aim to improve the model's accuracy and efficiency through further training and tweaking of parameters.

3. **Comprehensive Testing and Evaluation**: Once all components of the project are complete, we will carry out comprehensive testing and evaluation to ensure the system's effectiveness and reliability. This will involve testing with varied resume datasets, evaluating the accuracy of classification and ranking, and assessing the system's performance in real-world scenarios.

4. **System Integration**: Finally, we aim to integrate the parsing, classification, and ranking components into a unified system that can efficiently process resumes and output a shortlist of candidates for a specific job opening. This unified system will represent the final product of our project.

Moving forward, our focus will be on addressing the pending objectives, continually refining our approaches based on testing results and feedback, and ensuring the successful completion of the project.

REFERENCES

[1] *L. Zhang, "A Deep Learning Approach for Resume Information Extraction," in IEEE Access, vol. 8, pp. 22248-22258, 2020.*

[2] *K. Dahiya and D. Singh, "Resume ranking using natural language processing and clustering," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 133-138.*

[3] M. D. Ekstrand, M. Ludwig, J. A. Konstan and J. T. Riedl, "Rethinking the recommender research ecosystem: Reproducibility, openness, and LensKit," in Proceedings of the fifth ACM conference on Recommender systems, pp. 133-140, 2011.

[4] [4] Vaswani, A., et al. "Attention is All You Need." 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[5] S. S. Rautaray and A. Agrawal, "Resume Parser with Named Entity Recognition," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Bhubaneswar, India, 2020, pp. 1-4.

[6] Vaswani, A., et al. "Attention is All You Need." 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[7] [3] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in AAAI'06: Proceedings of the 21st national conference on Artificial intelligence, July 2006.