Natural Language Processing

Project 1

Done By:

## 

# Abstract

This project presents a solution for optimizing the matching between resumes (CVs) and job descriptions (JDs) using advanced natural language processing (NLP) and machine learning techniques. The proposed methodology involves collecting and preprocessing CV and JD datasets, including steps like stopword removal, Named Entity Recognition (NER), and Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction. By leveraging contrastive learning and embedding generation, the system computes precise matching scores between CVs and JDs, enhancing the recruitment process. The approach ensures a meaningful alignment of candidate profiles with job requirements, leading to more effective job recommendations. This work offers a structured approach to improving the accuracy of automated job matching, providing a foundation for future advancements in this domain.

# Credits

We would like to acknowledge the following individuals and resources for their contributions to this project:

\begin{itemize}

\item \textbf{Anna Wroblewska}, for his invaluable guidance and support throughout the project.

\item \textbf{The authors of the mentioned papers}, whose methodologies and insights were integral to shaping our approach.

\end{itemize}

We also wish to thank our peers and tutors for their constructive feedback and contributions to the project.

# Introduction

Natural Language Processing (NLP) is a field of artificial intelligence (AI) that enables computers to understand, analyze, and generate human language. It combines techniques from computer science, linguistics, and statistics to process large amounts of textual or spoken data. NLP applications are vast and include chatbots, sentiment analysis, image captioning, and speech recognition. The main goal of NLP is to make human-machine interactions more natural and intuitive.In addition, NLP can be extremely useful in helping with day-to-day tasks such as recruitment.

The objective of the project is to create a tool which can extract relevant information from CV´s which can be used for the human resources departments to classify candidates for initial recruitment phases,which usually are online. Also, it can be used for candidates to have job recommendations. NLP techniques will be used to identify the relevant information of CV’s like years of experience, skills and location.

In this document, we will begin by conducting a comprehensive literature review of state-of-the-art (SOTA) methods and tools in NLP, recommendation systems, and CV-JD matching. This review will explore existing research and methodologies, with a particular focus on comparison tables for datasets, methods, and pre-trained models. Following this , we will develop and document the solution concept for feature extraction from CVs and JDs, focusing on the project’s approach and methodology. This will include a detailed explanation of the approach, covering methods for feature extraction, merging, and initial model selection.

# Related work

### Project overview

# Little resume of the Literature Review & Background

**Resume Parser Using Natural Language**

**Processing Techniques**

The paper aims to streamline the recruitment process by developing an automated system that parses, structures, and ranks resumes based on job requirements using Natural Language Processing (NLP) techniques. The proposed solution involves a job portal where job seekers upload their resumes, which are then converted into plain text using Optical Character Recognition (OCR).

Key NLP techniques, including lexical analysis, syntactic analysis, semantic analysis, and Named Entity Recognition (NER), are utilized to extract and structure relevant information such as education, experience, and skills. The system also incorporates social media data for comprehensive profiling and ranks candidates objectively using ElasticSearch-based dashboards. This approach enhances recruitment efficiency while reducing biases and effort for both recruiters and candidates.

**A CV Parser Model using Entity Extraction Process and Big Data Tools**

This paper discusses the implementation of a resume parser using nlp,but taking into account the problem that CV´s are usually unstructured data and are not stored in conventional databases.The authors explore how to extract appropriately the entities of the resumes and how to optimize the process using big data tools.

This project also uses NER technique to identify the entities of the CVs and more advanced techniques such as entity relation extraction which connects relationships shared between different entities. It also reveals any conne-tions or events shared among these entities. The main aspect is it also establishes the indirect relationships amongst indirect connections.It also mentions common NLP techniques such as pos tagging or tokenization for the CV reusmer

The main goal is to extract the necessary entities from resumes ,which can be in a lot of different formats,such as XMl.In this step, the main tool used in NER.With this process,they can extract the main features for a lot of different types of inputs.However,they faced with the problem that words can have different meanings depending on the context.FOr this reason,more advanced techniques as linking or entity relation extraction.For optimizing the process of dealing with a huge amount of data they use Hadoop mapReduce to process the words in parallel.

**CONFIT: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning**

The paper presents C ON F IT, a simple and general-purpose approach to model resume-job matching using contrastive learning and data augmentation. To address the label sparsity problem in person-job fit datasets, C ON F IT first uses data augmentation techniques to increase the number of training samples. It employs two paraphrasing methods - EDA (Easy Data Augmentation) and ChatGPT - to create augmented versions of resumes and job posts.

The first problem they encountered is Resume-job datasets often suffer from sparsity of interaction records, as job seekers only apply to a few jobs. CONFIT addresses this by using data augmentation techniques like paraphrasing to create more training samples. . CONFIT then uses contrastive learning to further increase the number of training samples, going from B pairs per batch to O(B^2) pairs. This improves the quality of the learned resume and job embeddings.

The system's ability to efficiently rank large numbers of resumes and jobs is highlighted, making it a robust solution for person-job fit systems. Ethical considerations regarding biases and privacy are discussed, emphasizing the importance of mitigating biases in such systems.

The performance is demonstrated on the experiments on two real-world person-job fit datasets ,which show that C ON F IT outperforms prior state-of-the-art methods by up to 19% and 31% absolute in nDCG@10 for ranking jobs and ranking resumes, respectively. It also remains competitive on classification tasks despite not directly optimizing for them. The authors find that the data augmentation and contrastive learning components are crucial to C ON F IT's strong performance.

### Background

# Solution Concept & Approach

Now that we've done a little review of the Literature Review & Background we are going to see the Solution Concept & Approach.

| **Paper Title** | **Techniques/Tools** | **Objectives** | **Challenges** |
| --- | --- | --- | --- |
| Resume Parser Using Natural Language Processing Techniques | OCR, NLP techniques (lexical analysis, syntactic analysis, semantic analysis, NER), ElasticSearch dashboards | Streamline recruitment by parsing, structuring, and ranking resumes based on job requirements. | Handling biases, integrating social media data, and ensuring efficient information extraction. |
| A CV Parser Model using Entity Extraction Process and Big Data Tools | NER, entity relation extraction, tokenization, POS tagging, Hadoop MapReduce | Extract entities from resumes in varied formats and optimize processing with big data tools. | Managing context-dependent meanings of words and handling diverse data formats. |
| CONFIT: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning | Data augmentation (EDA, ChatGPT), contrastive learning, embedding ranking | Enhance resume-job matching by addressing label sparsity and improving embedding quality. | Overcoming dataset sparsity, mitigating biases, and ensuring ethical use of data. |

The proposed solution aims to optimize the matching between resumes (CVs) and job descriptions (JDs) through a structured approach. The first step involves data collection. The necessary datasets include CVs, which contain candidate profiles with information such as skills, experience, education, and job history, and JDs, which outline the job requirements, responsibilities, and desired skills.

Once the datasets are collected, the next phase involves data cleaning and preprocessing, which is critical for ensuring the accuracy and relevance of the analysis. This process begins with stopword removal, which eliminates common, non-informative words such as "am," "is," "are," "of," "a," and "the." These words do not contribute meaningful information to the underlying topics and are removed to enhance the dataset's semantic clarity. Next, non-English rows are discarded to ensure linguistic consistency across the dataset, which is essential for accurate natural language processing (NLP).

Following this, Named Entity Recognition (NER) is employed using SpaCy to extract meaningful entities from both CVs and job descriptions (JDs). Key focus areas for extraction include job titles, skills, certifications (e.g., programming languages, tools), as well as education and years of experience, which are crucial for aligning candidate qualifications with job requirements.

In addition, high-frequency words that appear in a significant percentage of documents (e.g., 80–90%) are removed, as these terms are often non-distinct and fail to provide valuable differentiation between various job roles or qualifications. These preprocessing steps align with methodologies proposed by Jagwani et al. (2023), who demonstrate the effectiveness of combining entity detection using SpaCy's NER and Latent Dirichlet Allocation (LDA) to extract meaningful semantic representations for resume evaluation. Their approach focuses on creating a content-driven scoring system, achieving an overall accuracy of 82% by considering attributes such as education, work experience, and skills (Jagwani et al., 2023).

Our proposed approach incorporates Term Frequency-Inverse Document Frequency (TF-IDF) as a key technique for feature extraction from resumes (CVs) and job descriptions (JDs). TF-IDF is widely used in information retrieval and natural language processing to identify and prioritize words or phrases that are not only frequent within a document but also distinctive across a corpus. By applying TF-IDF, the system ensures that terms specific to a resume or job description—such as technical skills, certifications, or unique job requirements—are given higher weight, while common words that appear across documents are deprioritized.

TF-IDF operates by calculating two key components:

1. Term Frequency (TF): This measures how often a term appears in a document relative to the total number of terms in that document.
2. Inverse Document Frequency (IDF): This scales the importance of a term inversely with its frequency across all documents in the corpus. Words that are common across many documents, such as "and" or "the," receive lower scores, while unique terms are weighted more heavily.

Where:

By leveraging TF-IDF, the proposed system extracts key features from CVs and JDs that are particularly relevant to matching candidates to job postings. This approach aligns with the methodologies discussed in Das et al. (2018), where the authors emphasize the importance of feature extraction from unstructured data for better alignment with recruitment processes. The TF-IDF model not only captures domain-specific terminology but also reduces the influence of generic terms, enabling a more precise and meaningful representation of resumes and descriptions.

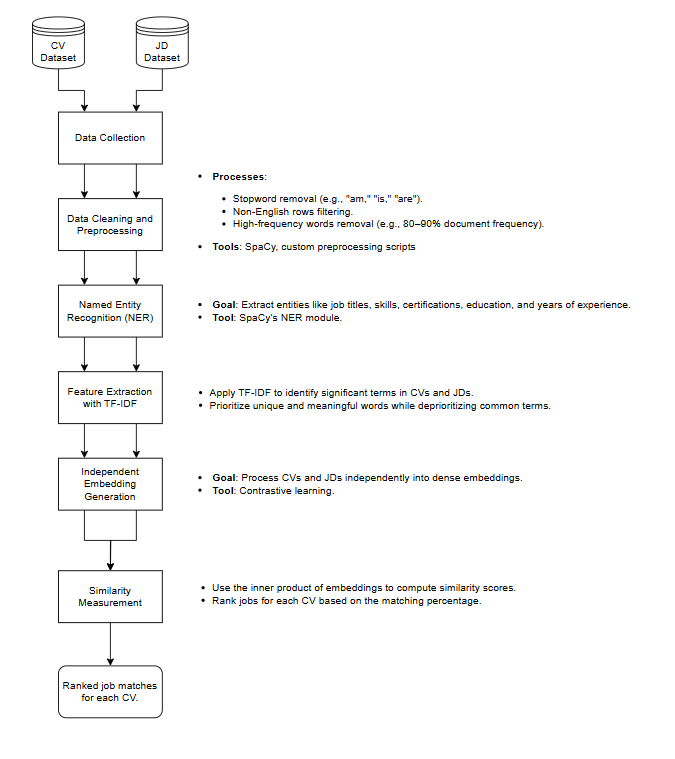
This feature extraction forms the foundation for subsequent processes in the pipeline, such as embedding generation and similarity computation, ensuring the system focuses on critical aspects of the resume-job matching task.

In order to maintain clarity and avoid confusion, CV and JD datasets will be processed separately. This independent preprocessing ensures that the unique attributes of each dataset are preserved, enabling the model to effectively differentiate between the qualifications presented in CVs and the requirements outlined in JDs. Specifically, CVs capture information such as experiences, skills, and educational qualifications, while JDs emphasize job-specific requirements, responsibilities, and desired expertise. By processing these datasets independently, the model can accurately represent the distinct characteristics of each, leading to more effective resume-job matching.

This methodology aligns with the approach described in *ConFit: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning* by Yu et al. (2024). The authors emphasize the importance of representing CVs and JDs as dense embeddings to encapsulate their unique information and improve matching accuracy. Their work demonstrates that independent preprocessing and representation of these datasets enhance the system’s ability to rank jobs and resumes effectively, achieving significant improvements in performance metrics such as nDCG@10 (Yu et al., 2024).

Contrastive learning, a machine learning technique, is particularly effective in representation learning. By training the model to distinguish between similar and dissimilar data pairs, it ensures that similar items are closely clustered in the embedding space while dissimilar ones are positioned further apart. This approach is also used in the methodology described in *ConFit: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning*, where contrastive learning is leveraged to address data sparsity and improve matching accuracy. By embedding resumes and job descriptions into a shared vector space, the model enables efficient similarity measurement, achieving significant improvements in ranking performance over traditional approaches such as BM25 and neural embeddings. The use of inner product similarity to compute matching scores further demonstrates the effectiveness of this strategy, which serves as a basis for our implementation plan (Yu et al., 2024).

To improve the accuracy and relevance of resume-job matching, we will employ contrastive learning as a core technique for embedding generation and similarity measurement. This approach enables the system to learn dense embeddings for both CVs and JDs within a shared vector space, facilitating a more precise computation of matching scores. The inner product of these embeddings will then be used to calculate a matching score, ranking jobs for a CV by their matching percentage. This methodology ensures that resumes and job descriptions are not only compared on textual similarity but also on contextual alignment.



"Pipeline Diagram for the Proposed Resume-Job Matching System: This flow illustrates the sequential steps involved, starting from data collection of CVs and JDs, through preprocessing, entity extraction, feature extraction using TF-IDF, embedding generation with contrastive learning, and similarity computation, leading to ranked job matches for each CV."

# Conclusion

In this document, we explored the potential of Natural Language Processing (NLP) to optimize the recruitment process by developing a system that efficiently matches resumes (CVs) with job descriptions (JDs). The project aimed to address key challenges in recruitment, such as unstructured data, large-scale datasets, and the need for accurate candidate-job matching, by leveraging state-of-the-art NLP techniques and methodologies.

Through a comprehensive literature review, we examined existing works and methodologies, including Named Entity Recognition (NER), Term Frequency-Inverse Document Frequency (TF-IDF), and contrastive learning, which informed the foundation of our proposed solution. This approach integrates multiple preprocessing techniques, feature extraction methods, and advanced machine learning algorithms to process CVs and JDs independently while preserving their unique attributes.

Our proposed system demonstrates the ability to extract critical information such as job titles, skills, certifications, and years of experience from CVs and JDs, using tools like SpaCy and TF-IDF. The solution further enhances the matching process by generating dense embeddings through contrastive learning, enabling precise similarity computations and ranked job recommendations for each candidate.

By combining insights from the literature with a structured, modular approach, the project offers a scalable, automated solution that reduces the time and effort required in the recruitment process for both candidates and human resources teams. Additionally, the methodology ensures a fairer and more efficient recruitment process by aligning candidate profiles with job requirements based on contextual and semantic relevance.

This work highlights the importance of NLP in transforming recruitment and sets the stage for further advancements, such as incorporating multilingual datasets, improving bias mitigation, and integrating more dynamic models for real-time job matching. As a result, the project not only contributes to improving recruitment efficiency but also paves the way for innovative applications of NLP in human resources and beyond.

The next phase of the project will involve creating a Reproducible Proof of Concept, including Exploratory Data Analysis and preliminary machine learning models. The focus will be on ensuring reproducibility, with source code, pre-processed datasets, and model parameters shared via a merge request.

# References

Das, P., Pandey, M., & Rautaray, S. S. (2018). *A CV Parser Model using Entity Extraction Process and Big Data Tools*. *International Journal of Information Technology and Computer Science*, 10(9), 21-31. <https://doi.org/10.5815/ijitcs.2018.09.03>

Jagwani, V., Meghani, S., Pai, K., & Dhage, S. (2023). *Resume Evaluation through Latent Dirichlet Allocation and Natural Language Processing for Effective Candidate Selection*. arXiv.<https://doi.org/10.48550/arXiv.2307.15752>

Yu, X., Zhang, J., & Yu, Z. (2024). *ConFit: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning*. arXiv.<https://doi.org/10.48550/arXiv.2401.16349>