Justin Magalona, Kalyaan Narnamalpuram, Yash Sonpal
Group 9
IE7500

Automated Resume Classification and Job Matching Using NLP Techniques

Project Description:
The goal of this project is to create an automated system which classifies resumes into different job domains and matches them with relevant job postings using Natural Language Processing (NLP). The project applies cutting edge methodologies like word embeddings, LSTM networks and Transformer models to process and understand unstructured texts found in resumes and job descriptions. The goal of the system is to help recruiters efficiently filter candidates while assisting applicants in finding suitable positions that precisely match their qualifications, which solves a crucial problem in contemporary digital hiring systems.

Problem Statement:
The manual effort involved in screening resumes and matching them with job postings is highly monotonous and prone to significant lapses in accuracy. On one hand, recruiters have to deal with the tedious task of screening through huge volumes of resumes, and on the other side, most of the applicants are unable to pinpoint the exact job openings that would be ideal for their skills and experience. This project proposes a novel solution to such inefficiencies through the application of NLP techniques to text analysis, categorization, and semantic matching of resumes with job descriptions. NLP, in this instance, is applied to enable dramatic improvements in the recruitment process through turning unstructured language data into valuable insights.

Data Sources:
Arsh Koneru. (n.d.). Linkedin Job Postings. Retrieved May 23, 2025 from
https://www.kaggle.com/datasets/arshkon/linkedin-job-postings

Snehaan Bhawal. (n.d.). Resume Dataset. Retrieved May 23, 2025 from
https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset/data

Data Description:

The Resume Dataset contains unstructured text data from resumes from different job applicants. It contains the text content of the resume, formatted HTML, and the job category. The dataset contains over 2400 resumes.

Column	Description
ID	Unique identifier for each resume
Resume_str	Text version of the resume in string format. Resume includes experience, skills, and education.
Resume_html	HTML formatted version of resume. Preserves original format of the resume. Includes lists,

	headings, etc.
Category	Job category for the resume (HR, Designer, IT, Automobile, etc.

The LinkedIn Job Postings dataset contains data about job listings that were taken from LinkedIn. The dataset includes job descriptions, job categories, and salary information. The dataset contains over 124,000 job postings.

Column	Description
job_id	Unique identifier for each job posting
company_id	Unique identifier for posting company
title	Job title
description	Job description
max_salary	Max salary
med_salary	Median Salary
min_salary	Minimum Salary
pay_period	Pay period for salary (Hourly,Monthly,Yearly)
formatted_work_type	Type of Work (Full time, part time, contract)
location	Job Location
applies	Number of applications submitted to posting
original_listed_time	Original time job was listed
remote_allowed	If job allows remote work
views	Number of times job has been viewd
job_posting_url	URL to the job posting on a platform
application_url	URL where application can be submitted
application_type	Type of application process (offsite, onsite)
expiry	Expiration date of job listing

closed_time	Time to close job listing
formatted_experience_level	Job experience level (entry, associate, executive)
skills_desc	Description detailing required skills for job
listed_time	Time when job was listed
posting_domain	Website domain with application
sponsored	Whether job is sponsored or promoted
work_type	Type of work associated with job
currency	Kind of currency in which the salary is provided
compensation_type	Type of compensation for the job

Expected Outcomes

Our end goal of this project is to create a rudimentary system that automatically reads and understands resume content and subsequently suggests under which category of job it best fits. For example, if one uploads a resume with programming experience, it should recognize it as best suited for an Information Technology job. Our goal is to make it easier so that recruiters can quickly sort through multiple resumes and job applicants can get an idea of what type of employment they might be qualified for. We will focus on categorizing the resumes by HR, Finance, IT, etc., based on the typed information in each resume in the first half of the project. We will be using a dataset that has already been preloaded with sample resumes belonging to familiar job categories, on which we will train our system to predict new unseen resumes. During the latter half, we will try to match these resumes to job postings on job sites like LinkedIn. So we'll analyze what's included in a resume, contrast it with text in job ads and decide which jobs are most similar or suitable. By the end of the project, we want to have a functional model that can take resume text as input and suggest both a job category and potentially similar job ads. This will give us a practical understanding of the role natural language processing can play in actual hiring systems, and we believe this project is a worthwhile effort to learn about how text data can be leveraged to make more informed decisions.

To ensure that our model is performing as expected, we will check if it can indeed predict the work category of the resumes that it has not encountered previously. We will do this by splitting our dataset into two partitions: training and testing. After having trained the model, we will then give the model a resume from the test set and check what job category it predicts against the actual job category in the data. To measure how accurate the model i.e. how often is the model's prediction right. We will observe some straightforward metrics such as precision and recall, which can tell us how well the model can select specific categories. Then, if we include job matching, we will check if the jobs that the model has proposed are consistent with the resume through real life instances and check if they make sense.