Empowering Job Seekers with Advanced NLP: A Revolutionary Approach to Career Navigation in the Modern Job Market

Class: DATS 6202– Natural Language Processing

Team 2: Amjad Altuwayjiri, Muhannad Alwhebie, Nammin Woo

I. Introduction

- This project addresses the challenges faced by job seekers in navigating the job market by providing a user-friendly tool that leverages advanced NLP techniques to connect them with career opportunities that align with their qualifications and aspirations, thereby streamlining and enhancing their job-hunting experience.

I. Introduction

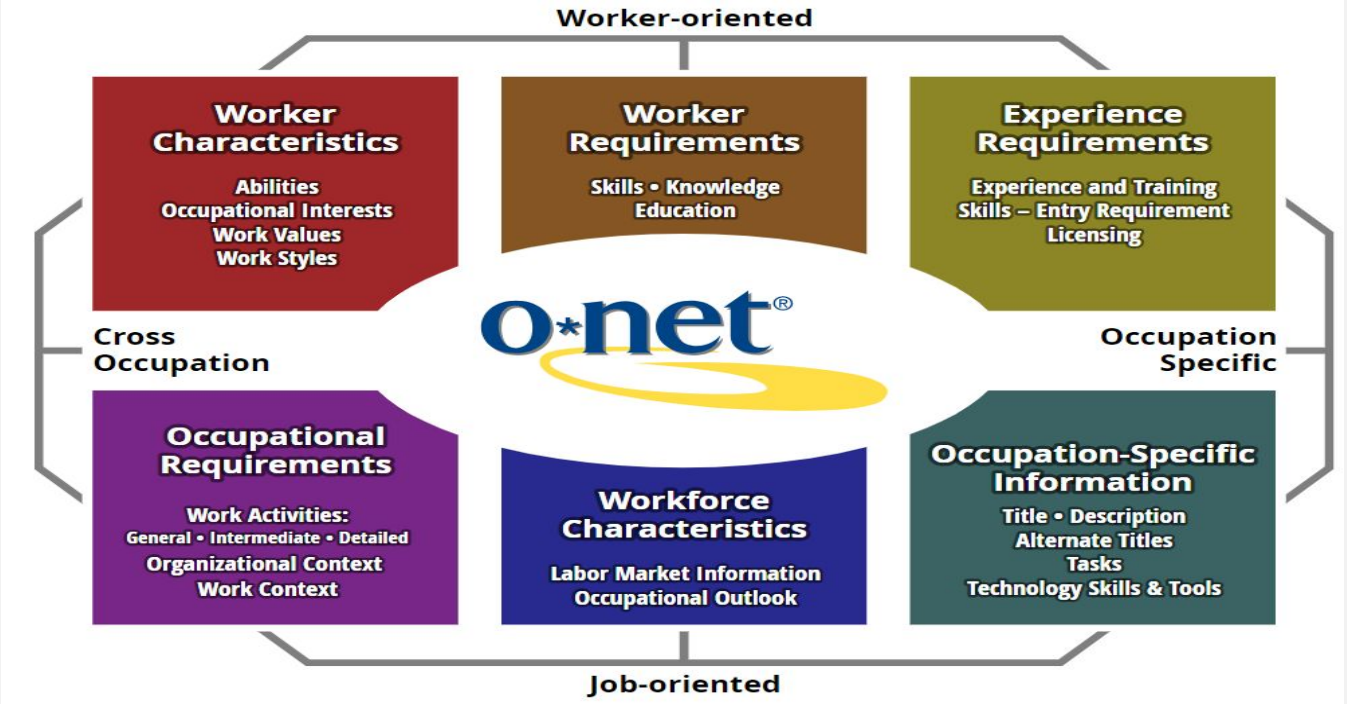
- The project has two main parts:
- 1 Text Classification Model for predicting job levels in resumes
  2. Semantic Similarity Search using Siamese-BERT Networks.

Data

Classification

Semantic similarity

- ONET® Database by the U.S. Department of Labor
- Rich data set with detailed information, including job responsibilities, required skills, and educational requirements



Data Processing Workflow

Classification

Semantic similarity

Step 1: Data Preparation

- ❏ Unzip Bulk Data: Process a .zip file containing 40 individual text files.
- ❏ Content Types: Files include Base information (code-description mapping) and Element Contents in various subcategories.

Step 2: Corpus Generation

- ❏ Select Key Categories: Focus on 5 Resume-related categories - Ability, Skill, Tech-Skill, Knowledge, Interest.
- ❏ Primary Aggregation: Combine Contents data with Base information.
- ❏ Create Job Corpus: Extract 5 highest demand elements from about 50 unique elements in each category.

Step 3: Corpus Aggregation and Classification

- ❏ Generate Full Corpus: Combine Summarized job descriptions with job corpus from each category.
- ❏ Add Job Level Label: Prepare data for a classification model, labeling each job level.

Example of Individual corpus (Ability Category)

1. Abilities: Title-Element-Scale

- PK: SOC Code (873 Occupations) - Element (52 Abilities) - Scale ID (2 Types of value)
  - 90792 records (873 * 52 * 2)
  - Two Text descriptions: Description_SOC(Job), Description_Ele(Abilities)
  - Created 'Value_ratio': Data Value/Maximum
  - Element:
    - Name
    - Description (from Base information table)

  ex.Problem Sensitivity ability:The ability to tell when something is wrong or is likely to go wrong. It does not involve solving the problem, only recognizing that there is a problem.

Title	Element Name	Scale Name	Minimum	Maximum	Data Value	Value_ratio
Statisticians	Problem Sensitivity	Importance	1	5	3.38	0.676
Statisticians	Problem Sensitivity	Level	0	7	3.75	0.536

Data Processing Workflow

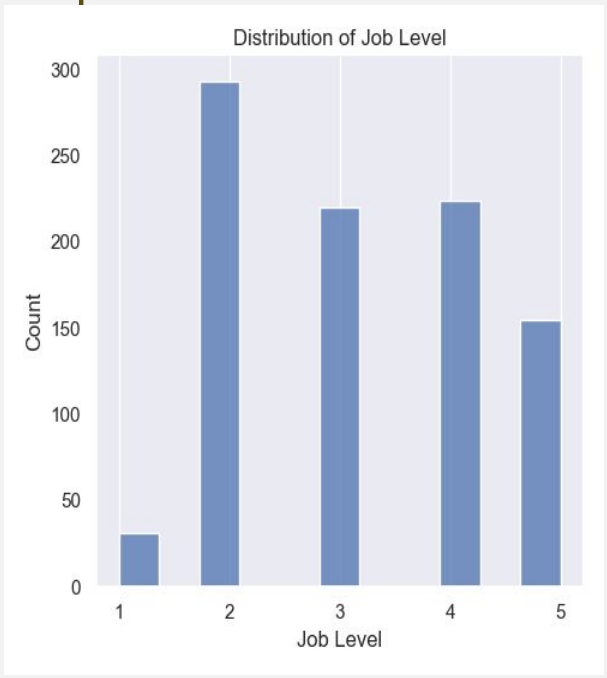
Classification

Semantic similarity

Final ONET Job Corpus Data has 1,016 records with 10 features. Job levels are uniformly distributed except level 1

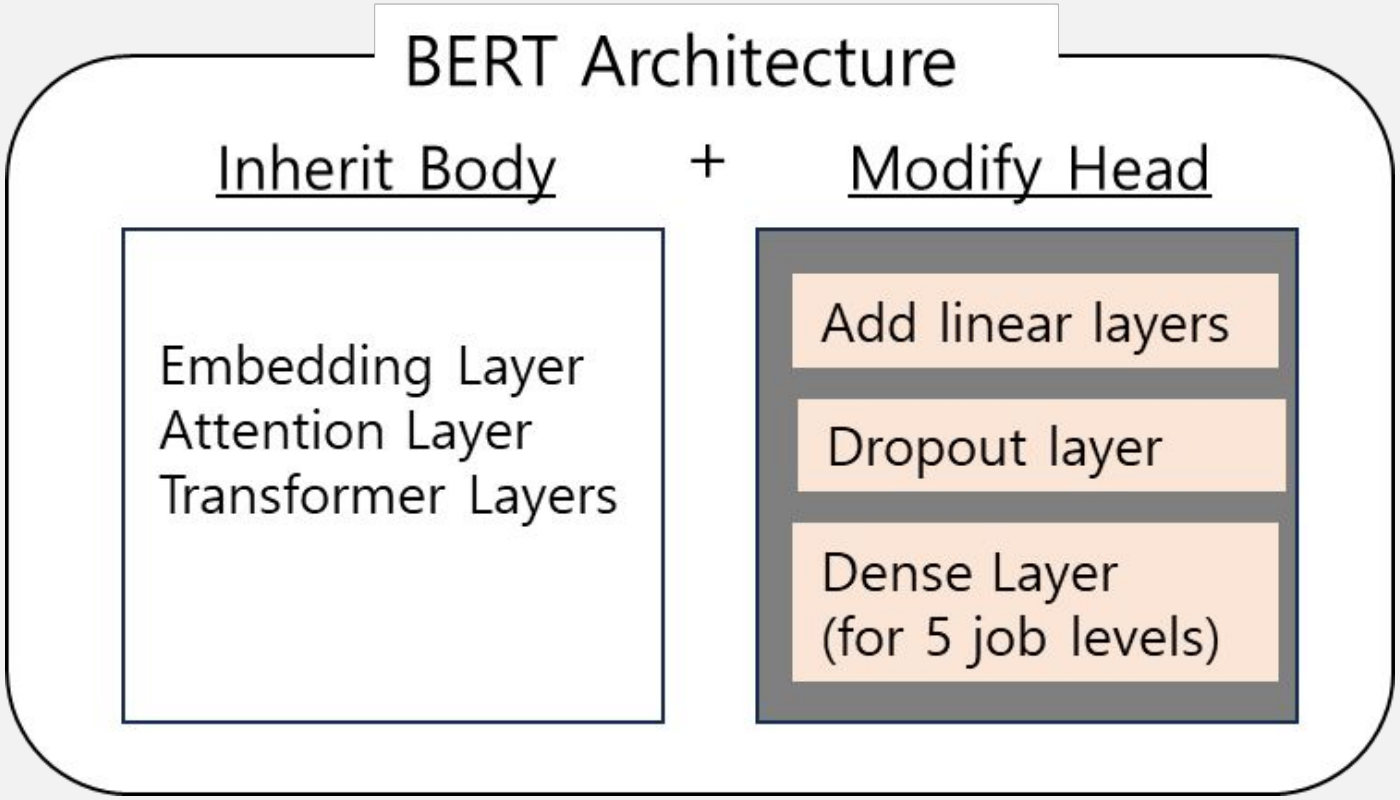
```
C.Features:
O*NET-SOC Code                 object
Title                          object
Description                    object
Description_Abilities          object
Description_Knowledge          object
Description_Skills             object
Description_Tech               object
Description_Interests          object
Description_Job                object
Job Zone                      float64
dtype: object
```

```
D.Null value check:
O*NET-SOC Code                 0
Title                          0
Description                    0
Description_Abilities          0
Description_Knowledge          0
Description_Skills             0
Description_Tech               0
Description_Interests          0
Description_Job                0
Job Zone                      93
dtype: int64
```



	Description_Job	Job Zone
Chief Executives	Determine and formulate polic...	5.0
Chief Sustainability Officers	Communicate and ...	5.0
General and Operations Managers	Plan, direct, ...	4.0

1. BERT transformer with a linear layer: 72%

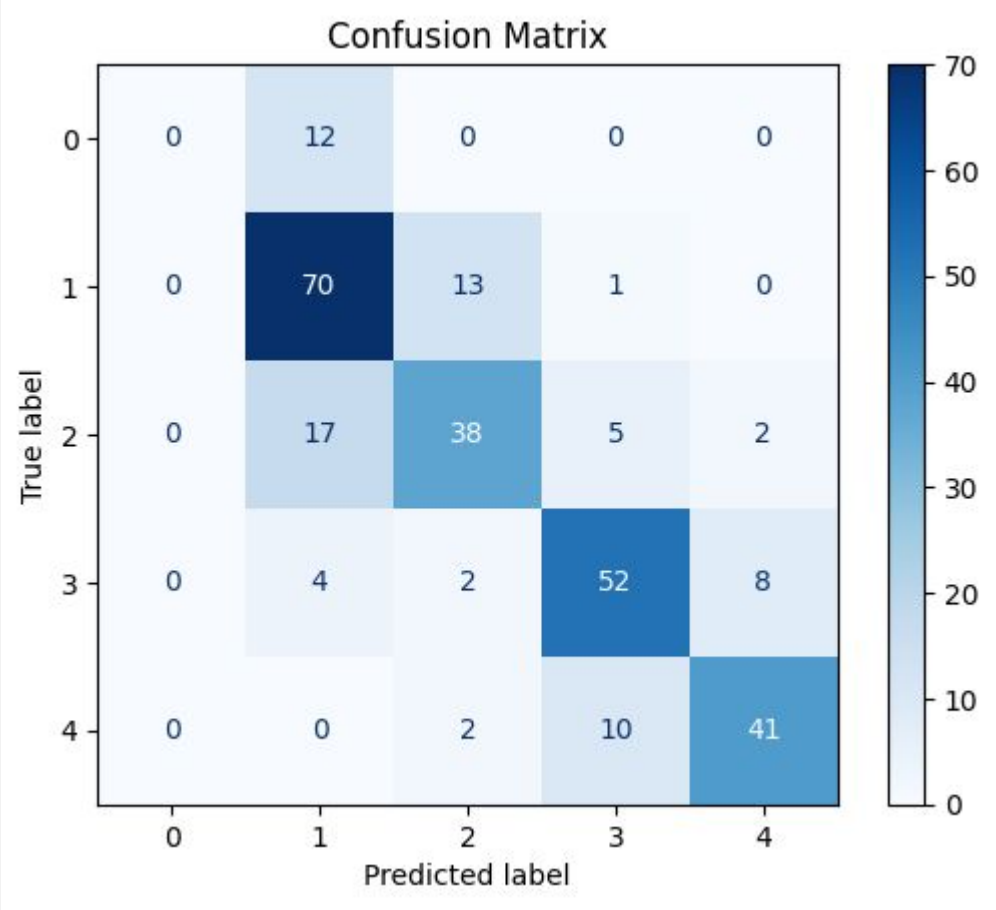


Data

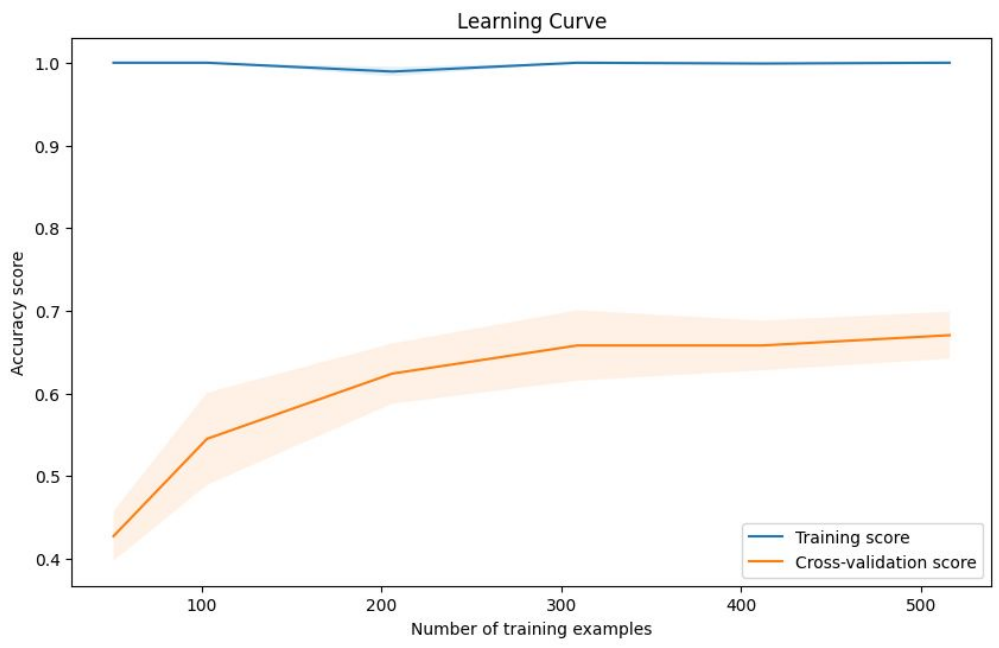
Classification

Semantic similarity

1. BERT transformer with a linear layer: 72%

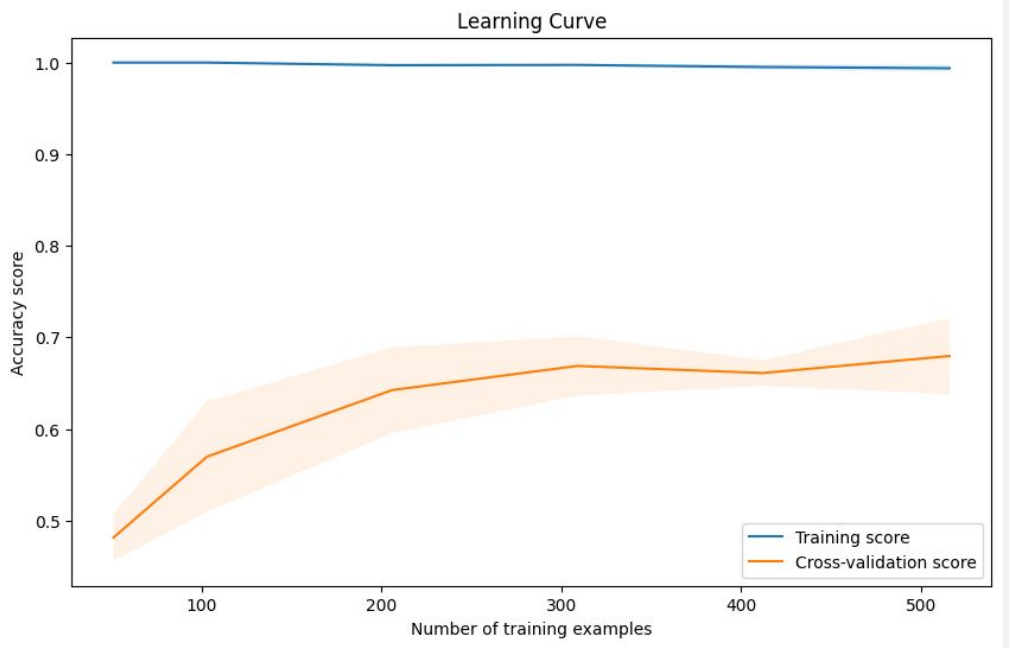


2. MLP



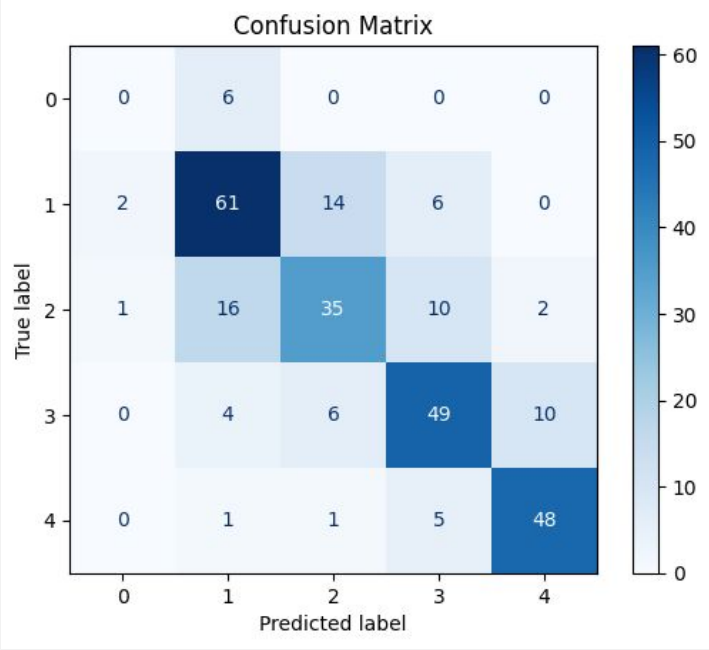
MLP

3. Logistic regression



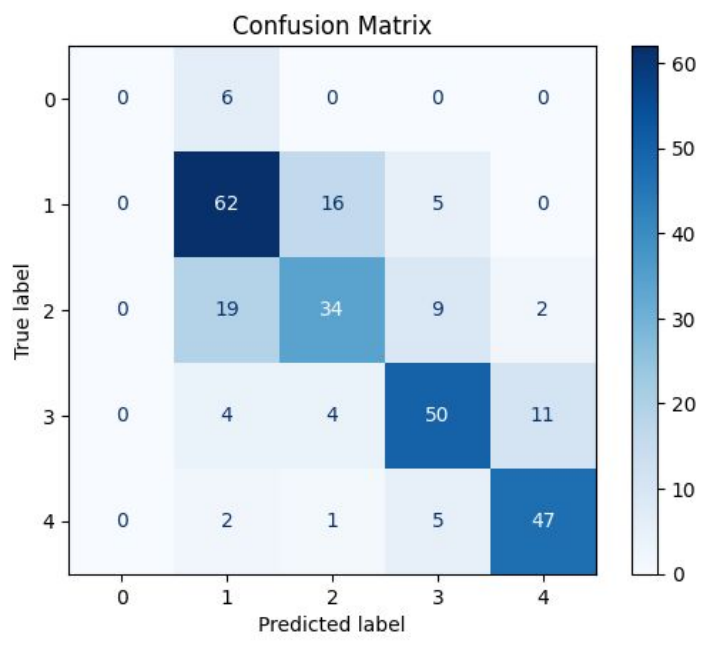
Logistic regression

2. MLP



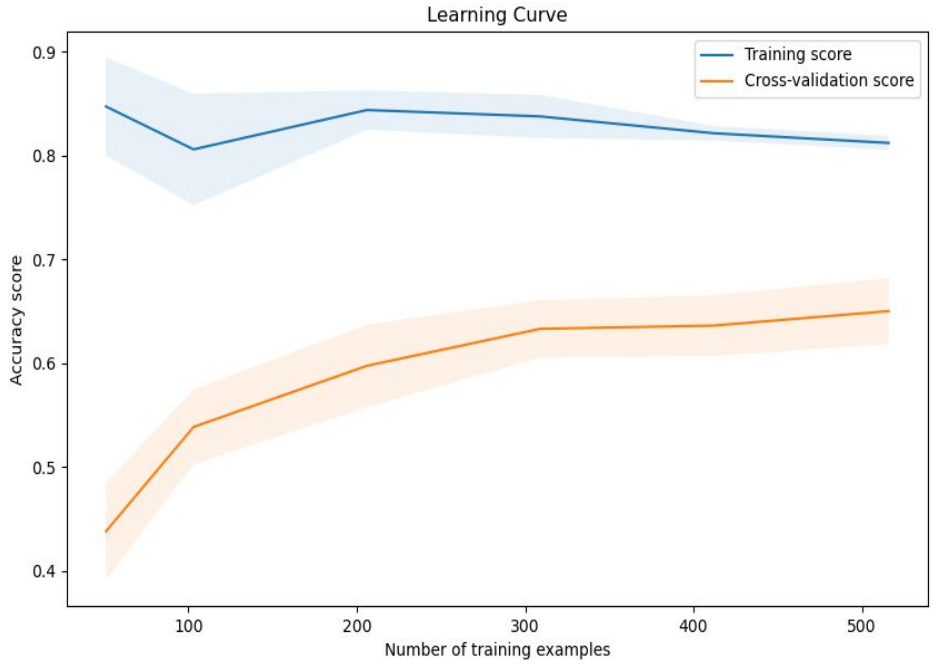
MLP

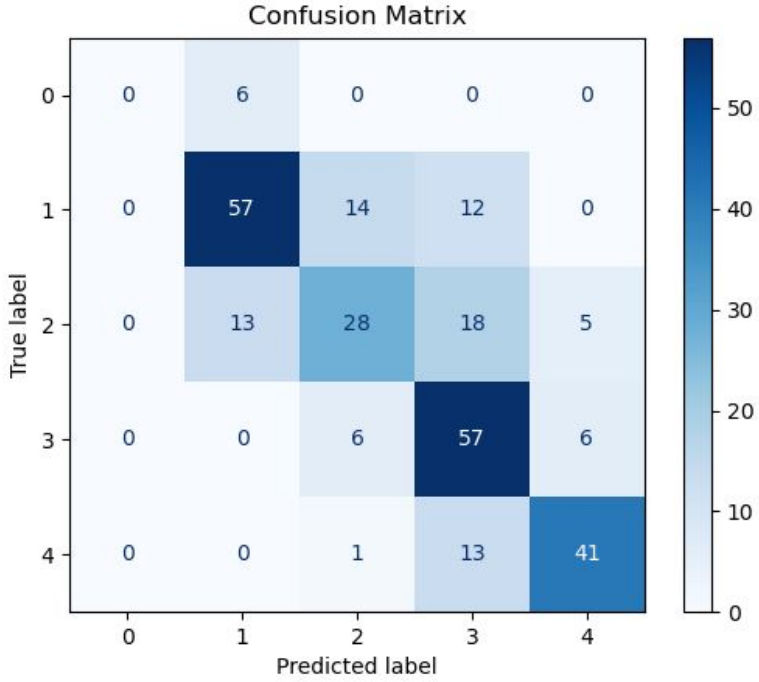
3. Logistic regression



Logistic regression

4. Naïve Base 66%





Data

Classification

Semantic Similarity

- Siamese-BERT Networks (Sentence-BERT)
- Semantic similarity searches
- Embedding vectors for both job description corpus and resumes

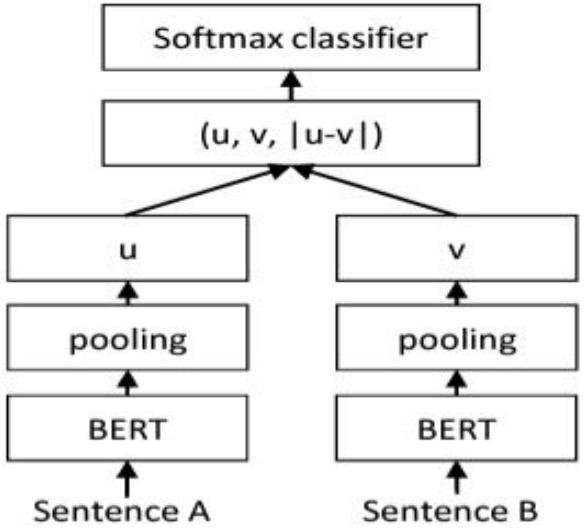


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

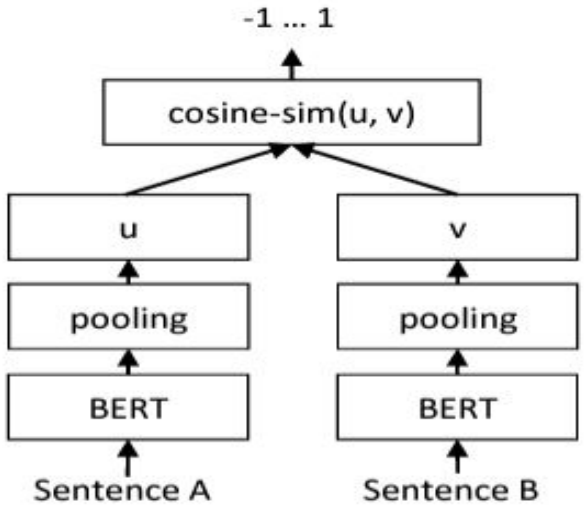
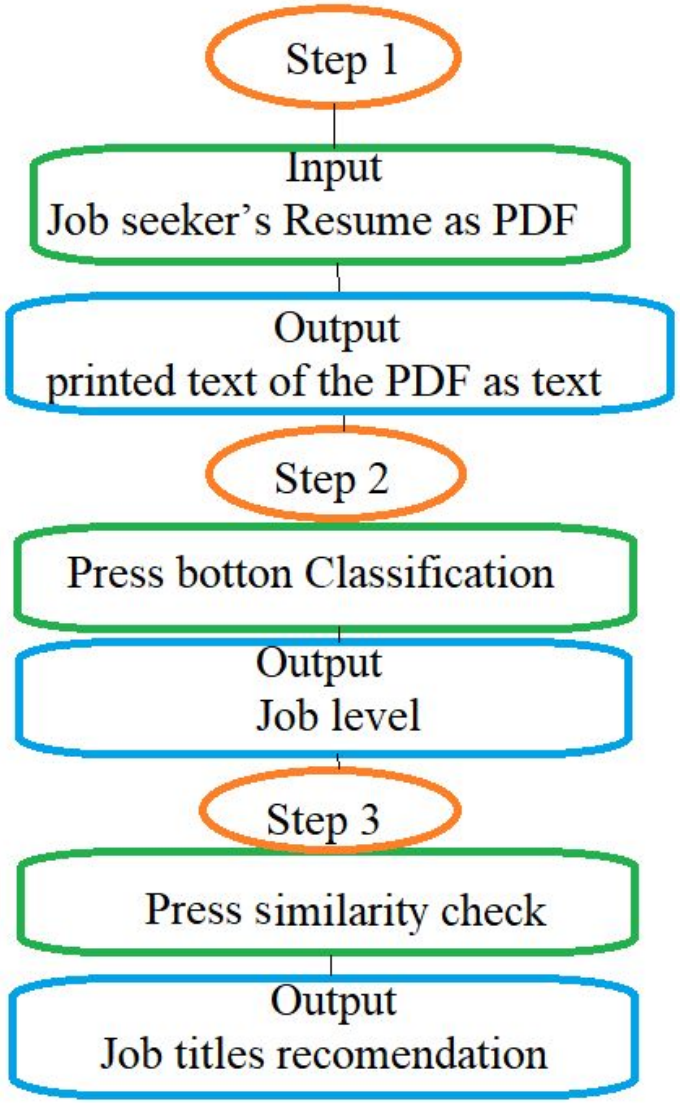


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Streamlit architecture



Application

Result

BERT got the highest accuracy of 72%, MLP and LG achieved accuracy of 69% , then Naïve Base of 66%.

Instead of personal guess of the job level and the title a job seeker looks for, we propose a systematic NLP approach by using a friendly user interface.

Expand the dataset to include a wider range of jobs and titles.

Thank you