Adaptable Framework to Automate the Classification of Occupation Codes

MINISTRY OF MANPOWER | MANPOWER RESEARCH & STATISTICS DEPARTMENT

Lucas Ng

2 December 2024

Agenda

- Objective
- Choice of Model
- End-to-End Modelling Workflow
- Model Results & Evaluation
- Use Case: Detection of new and emerging occupations from Online Job Vacancies
- Short Demo

Objective

Build a **text classification model** to assign the most appropriate **occupation code** based on open text info (e.g job title, job description).

Individual

- Job Title
- Job Description

Text classification model

Occupation Code

End-to-End model workflow

Overview

MINISTRY OF MANPOWER | MANPOWER RESEARCH & STATISTICS DEPARTMENT

Household/ Establishment Surveys

Non-synthetic & synthetic data

ChatGPT-40 mini

Custom Prompt

Data Generation

Synthetic Job Title | Synthetic Job Duties

Paraphrasing Step LLM: Gemma-2b (Fine-tuned via SFT)

Gemma Job Title

Gemma Job Duties

Occupation Definitions (i.e Job Title, Job Tasks and Duties

Model Training & Evaluation

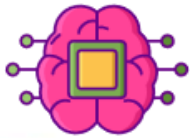
Synthetic Job Title	Gemma Job Title	Non-synthetic Job Title
Synthetic Job Duties	Gemma Job Duties	Non-synthetic Job Duties

f"job_title=<job_title> job_duties=<job_duties>"

f"job_title=<job_title>"

f"job_duties=<job_duties>"

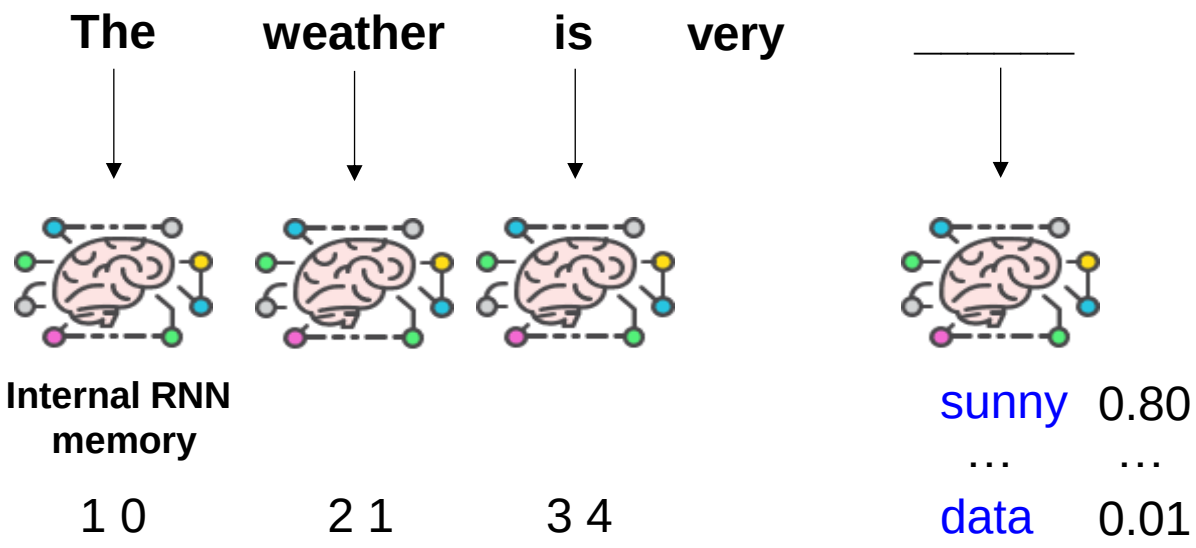
Embedding: BGE v1.5 Large



Neural Network

Choice of Model: Recurrent Neural Networks

Traditional recurrent neural networks (RNNs)



Challenges

- When trying to learn from or remember long pieces of information, **RNNs** may have trouble because they **lose or mismanage important details over time**.
- **Long short-term memory (LSTM)** networks have special features to overcome the issues faced by simpler models, allowing them to manage information more effectively. However, they may still suffer from the issues mentioned above.

Choice of Model: BERT

- Transformer-based language model
- Reads text bidirectionally, understanding context from both left and right
- Trained on Wikipedia and BooksCorpus
- Transformer architecture & attention mechanism

High attention **Low attention**

Sentence : The animal didn't cross the street because it was too tired.

End-to-End model workflow

Removing inconsistencies, mislabeled data

MINISTRY OF MANPOWER | MANPOWER RESEARCH & STATISTICS DEPARTMENT

Data Preprocessing | Synthetic data generation | Handling data imbalances | Stratified split

1 Data Preprocessing

Surveys

Challenges

- There may be **mislabeled** data points where the job title/descriptions do not correspond to the correct occupation code labels. This makes certain data points unsuitable for use to train the model, as the model may pick up wrong associations.
- **Impractical to manually vet** through large volumes of raw data to curate clean training data.

Solution

- A **Retrieval Augmented Generation (RAG)** solution was developed to automate the process of obtaining clean training data.
- **It compares the embedding of a** given job title and job duties against the entire list of occupation codes[1]. Returning the Top 5 best candidates.
- A question-answer prompt along with the candidates is passed to a **LLM** to select the best occupation code[2] along with their relevance scores.

[1]Details include a) Job Title, b) Job Duties, c) Examples of occupations with this occupation code.
E.g <job_title>Data Scientist</job_title><job_duties>Programming, Data Visualisation</job_duties><examples></examples>
[2] The LLM does not return a response if the job title/duties input lacks sufficient information/context for accurate prediction.

End-to-End model workflow

Automated labelling

MINISTRY OF MANPOWER | MANPOWER RESEARCH & STATISTICS DEPARTMENT

Data Preprocessing	Synthetic data generation	Handling data imbalances	Stratified split

Dataset with uncoded job title and job duties

Occupation Code book

Additional textual information for each occupation code

BGE v1.5 Large (Fine-tuned on question-context pairs) to enhance relevance in context matching, improving retrieval quality

Query embeddings

Vector store Index 1

Vector store Index 2

Qwen2.5

Qwen 2.5-7b Additional queries to enhance retrieval

Top 5 best candidates

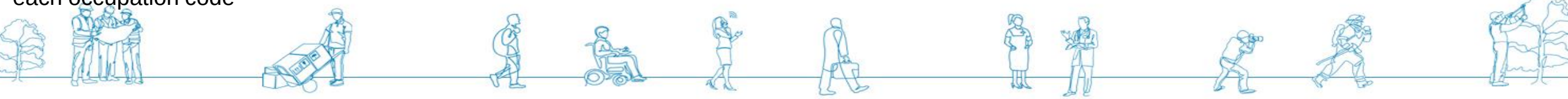
Top 5 best candidates

Qwen2.5

Qwen 2.5-7b (Fine-tuned on question-answer pairs)

Assign a relevance score to each candidate

Candidates with the highest relevance score is selected



End-to-End model workflow

2 Synthetic data generation

For each unique occupation code, we structure the job title, detailed definitions & examples of occupations as a custom prompt

#CONTEXT#
I would like to train a text classification model to predict an occupation code given a job title and job duties. As some labels have insufficient training examples, I require your help to generate synthetic data.
#OBJECTIVE#
Provide me with a list of 30 synthetic job titles and accompanying job tasks and duties.
#STYLE#
Follow the writing style of human resource and recruitment professionals.
#TONE#
Descriptive.
#AUDIENCE#
Job seekers.
#RESPONSE#
JSON output with {'job title':[job title 1, job title 2], 'job duties': [job duties 1, job duties 2]}

ChatGPT 40-mini

Synthetic data

End-to-End model workflow

3 Handling data imbalances

Challenges

- There are **imbalances** in the training data: certain occupation codes have many data points (**majority classes**) and others have only a few (**minority classes**).
- As a result, the training process would be dominated by the majority classes, potentially resulting in **poorer model performance for the minority classes**.

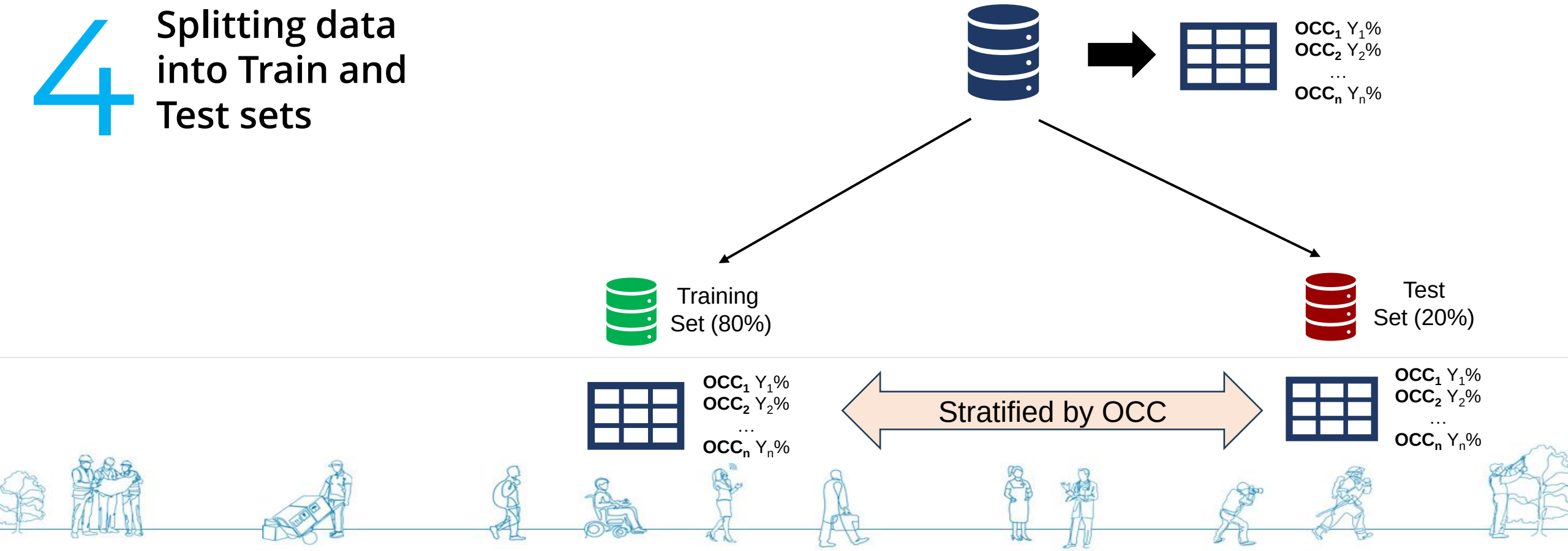
Solutions

Oversampling / Undersampling: **increase** / **reduce** the number of samples in the **minority / majority** classes to balance data.

Class-wise difficulty balanced (CDB) loss: CDB ensures that the model places greater emphasis on classes with weaker performance, thereby achieving a more balanced performance across both majority and minority classes.

End-to-End model workflow

4 Splitting data into Train and Test sets



Use Case: Detection of new and emerging occupations from Online Job Vacancies

MINISTRY OF MANPOWER | MANPOWER RESEARCH & STATISTICS DEPARTMENT

Online Job Vacancies

Original Job Title

Original Job Description

Occupation autocoder

- Custom neural network architecture
- Trained on synthetic and non-synthetic data

Each record assigned predicted;
1. Occupation code (Top N)
2. Confidence score (0 - 1)

Records with <=0.5 confidence for Top 1 occupation code

Qwen2.5

-Fine-tuned using SFT for Text Extraction & Skills Extraction

Original Job Title

Original Job Description

Skills

Qwen2.5

-Fine-tuned using SFT on question-answer pairs

NEO candidates

Human annotations (Feedback)

Reinforcement learning from Human Feedback (RLHF)

Try it out!

MINISTRY OF MANPOWER | MANPOWER RESEARCH & STATISTICS DEPARTMENT

Job Title

0/64

Job Description

Submit

MINISTRY OF MANPOWER

Thank you for your kind attention.