Project Outline for CSI-6900

Student Name: Pavan Balaji Kumar

Student Number: 300169572

Supervisor: Diana Inkpen

---

1 Introduction

The Curriculum Vitae's (CV) of researchers can be considered one of the important in the scientific community. The document containing information about the research projects, publications, awards, and conferences undertaken by them. It also provides plethora of other information such as organisation, supervisors, research standard and may more [1]. These are just some of the metrics which are used to evaluate a researcher performance in a particular discipline and academic success [2]. The CVs of a researchers also works as job search resource and are evaluated carefully before recruiting or collaborating with one another [2]. Though there is a standard for what information to include in the CV, there is no clear structure followed by the researchers, which makes the perusing and evaluating CV's a tedious process and a lot of research on automating this process is being done.

In this project we aim to automate the information extraction and evaluation of Scientific Curriculum vitae to increase the efficiency CV evaluation phase and help researcher's find suitable partners or assistant for their research work. This will be accomplished using various natural language processing techniques. Methods like Name entity recognition (NER), Part-of-Speech (POS) tagging will be used for extraction of required information from the CVs. Theses methods are trained to identify all the entities of the same type in a text or document efficiently [3]. Pre-trained deep learning models such as BERT, RoBERTa [4] will be used for the NER task to identify entities in the CVs like name publications, awards, conferences etc. This extracted information is processed into a Dataset for downstream tasks such as evaluation of an individual CV. This evaluation task can be a binary classification task or a summarisation task. This phase of the project will make use of deep learning models such as LSTM, Transformer models etc. and word embeddings to classify and summarise the information extracted from the Curriculum Vitae.

2 Project objectives

Since the we could not find an reliable dataset of Scientific CVs we are build our own dataset.

1. Collecting public CV available online and building a dataset consisting of adequate number of CVs. Tagging is done using online tools for the downstream task.

2. A baseline model is built using machine learning algorithms such as SVM, MLP etc. to be used for comparison with other models.

3. Advanced models like BERT, RoBERTa etc. are trained and NER is performed on the CVs and information is extracted.

4. The extracted information is processed into a dataset containing both textual and numerical features.

5. The dataset built is used to train deep learning models such as LSTM, Transformers and word embedding to perform classification or summarization of information extracted.

3 Project Deliverables

1. CV Dataset: Dataset built using Name entity Recognition method containing textual and numerical features.
2. NER model: Name entity recognition models used to extract information from CVs.
3. Deep learning model to classify or summarize the CV's information

4 Marking Scheme

The project will be marked in the following scheme:

1. 50% for implementation
2. 50% for report writing

5 Learning Objectives

The projects learning objective consists of the following topics:

1. Understanding and implementing various data preprocessing, data transformation and cleaning techniques.

2. Grasping the working of Name entity recognition process and implementing it to suit our project objectives.
3. Comprehension of various deep learning and machine learning algorithms like SVM, LSTM, RNN, Transformers, BERT, RoBERTa etc. and implanting them to accomplish our project goals.

6 Project tasks

The following are the main tasks:

1. Collecting public CV's and tagging them with entities to create a dataset for NER
2. Preprocessing the dataset by removing stop-words, using stemming and lemmatization, parsing and removing anomalies in the data
3. Training baseline model for comparison with others
4. Training model like BERT and RoBERTa to perform NER on CVs.
5. Evaluating the models and selecting the best model to perform NER.
6. Creating the Dataset from the Output of the NER model. Apart from features like name, publications, conferences etc. other feature are derived from them like number of publications, average impact factor etc. to help in the downstream task.
7. Preprocessing the textual data using word embeddings for the deep learning algorithms.
8. Training various deep learning models using LSTM, RNN, Transformers etc. to perform summarization and classification of CVs.
9. Evaluating the various models and evaluating the best model to complete the task.
10. Preparing a report summarizing the project.

References

[1] Cañibano, Carolina & Bozeman, Barry. (2009). Curriculum vitae method in science policy and research evaluation: The state-of-the-art. Research Evaluation - RES EVALUAT. 18. 86-94. 10.3152/095820209X441754.

[2] Fischer, J., Ritchie, E. G., & Hanspach, J. (2012). Academia's obsession with quantity. *Trends in ecology & evolution*, *27*(9), 473-474.

[3] Fareri, S., Melluso, N., Chiarello, F., & Fantoni, G. (2021). SkillNER: Mining and mapping soft skills from any text. *Expert Systems with Applications, 184* doi:10.1016/j.eswa.2021.115545

[4] Yang, S., Yoo, S., & Jeong, O. (2020). DeNERT-KG: Named entity and relation extraction model using DQN, knowledge graph, and BERT. Applied Sciences (Switzerland), 10(18) doi:10.3390/APP10186429