Resume Summarization using NLP

Imee Compra, Sophia Dalumpines, Christine delos Reyes, Lea Flor
College of Science, Technological University of the Philippines - Manila
imee.compra@tup.edu.ph
sophia.dalumpines@tup.edu.ph
christine.delosreyes@tup.edu.ph
lea.flor@tup.edu.ph

***Abstract* –** Resumes and Curricula vitae are important documents for proving one's eligibility for a job position. It serves as an overview of the background, skills, and achievements an applicant can offer. [1] After posting a job, employers face hundreds of resumes and applications and it is a hassle if they review them manually. Through the advancement of Artificial Intelligence (AI), Natural Language Processing (NLP) algorithms have been created. NLP has a lot of uses and an example of it is automatic summarization. NLP can be used to create a summary from a longer piece of text (which is commonly found in unstructured data). [2]

This project aims to produce a Resume and Curriculum Vitae summarizing tool with the use of a Natural Language Processing algorithm.

Keywords: Resume, Curriculum vitae, Artificial Intelligence, Natural Language Processing, and Named-Entity Recognition.

I. INTRODUCTION

Whether in large corporations or small businesses, the Human Resources department deals with hundreds of resumes every day. A resume is an essential component of an applicant's application. According to Forbes, the average number of applicants for one job opening is 118 [3]. This means that employers do not have enough time to thoroughly review each resume. Because there is no universal standard format or structure for resumes, it is more difficult for recruiters to pick out key points from an applicant's profile. To address this, this research focuses on a resume summarization model based on natural language processing.

Natural Language Processing (NLP), which focuses on teaching computers to understand text and words in the same way that humans do, is the most commonly used method for screening resumes [4]. The model will be able to extract key information from the resume using NLP. This includes the applicant's name, email address, degree, skills, and other vital information.

II. RELATED WORK

[5] NLP is a component of AI reasoning that handles analyzing, interpreting, identifying, and generating the dialects that people use in a distinctive way to enable interaction with PCs in both written and spoken contexts using typical human dialects rather than codes. Text summarization's goal is to extract important data from large amounts of information. The use of automated text summarization makes it simpler for users to extract crucial information from massive amounts of data. Other graph-based ranking systems include Positional Power Function, Hyperlink Induced Topic Search, and Text Rank. The Text Rank algorithm is a diagram-based text processing placement model that may be used to locate the most relevant sentences in a text as well as keywords that can be used for summarization.

[6] Large Corporate Companies and recruitment agencies receive and manage thousands of applications from job applicants. Those applications which usually consist of Resumes and Curriculum vitae are unstructured data that are automatically extracted from their systems. This unstructured data yields unexpected formats of information in the documents. Most companies rely on the Resume Ranking System and it's a Parser system that uses natural language processing (NLP) to parse the following candidate résumé and their social accounts. Without any manual intervention, that is.

One of the most common data preprocessing tasks is named entity recognition (NER). [7] Named-entity recognition (NER) is a preprocessing step of information extraction (IE) that seeks for and categorizes specific entities inside a text. Entity identification, chunking, and extraction are all terms used to describe NER. Natural language processing (NLP) is an application of NER in artificial intelligence (AI). NER systems have been developed that use textual grammar-based techniques as well as statistical models such as machine learning. This paper [8] improves their procedure by

summarizing and evaluating resumes based on matching the required skills data and the organization's cut-off. With the help of NER, resumes are collected in a directory and then run through the function call to extract the entities from the plain text. The entity extraction function is responsible for extracting entities from resumes by extracting tokens from the document content. Each resume's entities are saved in dictionary format, and the scoring is calculated based on the skill set provided by the employer.

The CV is an essential component of any interview. Their software [9] enables recruiters in screening resumes more efficiently, lowering recruiting costs. This will introduce the possible applications to the organization, and the candidate will be successfully put in a company that values his or her skill set and abilities. Their approach shortens the process by summarizing and classifying resumes based on how closely they match the organization's desired skills data and cut-off. This method evaluates candidates' abilities and ranks them based on skill criteria and the recruiting company's cut-off. Furthermore, a resume summary is provided for each candidate to provide a summary of the candidate's credentials.

[10] RoBERTa is a replicate of the BERT which trains a neural network model. RoBERTa can either match or exceed the performance of all the other BERT methods. The RoBERTa modifications include - lengthier, larger batch training of the model. Eliminating the next sentence prediction, longer sequences during training, and altering the training data's masking pattern in real time. RoBERTa also collects a larger amount of data.

III. EXPERIMENT

The researchers have employed Natural Language Processing to summarize resumes and curricula vitae (CVs). To achieve this goal, SpaCy, an open-source software library, was utilized to train a custom Named Entity Recognition (NER) model on a dataset comprised of 200 resumes and CVs.

A. Data Preparation

To prepare the data for training, the dataset was preprocessed by extracting the text and annotations in each document. This information was then stored in a DocBin object and saved to disk.

B. Named-Entity Recognition Model

Named-Entity Recognition (NER) is the process of locating and classifying the named entities from an unstructured text

into named entities (e.g., names, organizations, locations)

In the context of this project, a custom NER model was trained to parse and extract information from resumes and CVs. The model was trained using a configuration file and the aforementioned preprocessed dataset. Contained in the configuration file are the necessary hyperparameters to train the model such as the epochs and optimization steps.

C. *Testing*

To test the efficacy of the trained model, the researchers have used three resumes to test how well the model extracts and categorize the information from the resumes. The proposed system can extract information from documents, pdf, and text files.

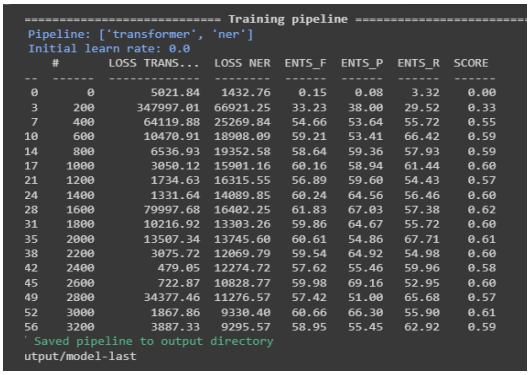
D. Limitations

Upon the completion of the testing period, the following limitations of the system were determined:

1. The system cannot perform well with resumes and CVs with complex designs.
2. The system cannot perform well with resumes and CVs containing images and diagrams.

IV. RESULTS

A. Training Result

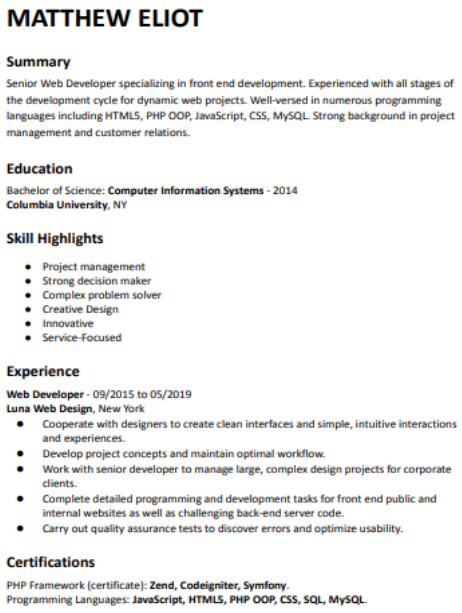


```
=========================== Training pipeline ===========================
Pipeline: ['transformer', 'ner']
Initial learn rate: 0.0
   #       LOSS TRANS...  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
--  ------  -------------  --------  ------  ------  ------  ------
 0       0        5021.84   1432.76    0.15    0.08    3.32    0.00
 3     200      347997.01  66921.25   33.23   38.00   29.52    0.33
 7     400       64119.88  25269.84   54.66   53.64   55.72    0.55
10     600       10470.91  18908.09   59.21   53.41   66.42    0.59
14     800        6536.93  19352.58   58.64   59.36   57.93    0.59
17    1000        3050.12  15901.16   60.16   58.94   61.44    0.60
21    1200        1734.63  16315.55   56.89   59.60   54.43    0.57
24    1400        1331.64  14089.85   60.24   64.56   56.46    0.60
28    1600       79997.68  16402.25   61.83   67.03   57.38    0.62
31    1800       10216.92  13303.26   59.86   64.67   55.72    0.60
35    2000       13507.34  13745.60   60.61   54.86   67.71    0.61
38    2200        3075.72  12069.79   59.54   64.92   54.98    0.60
42    2400         479.05  12274.72   57.62   55.46   59.96    0.58
45    2600         722.87  10828.77   59.98   69.16   52.95    0.60
49    2800       34377.46  11276.57   57.42   51.00   65.68    0.57
52    3000        1867.86   9330.40   60.66   66.30   55.90    0.61
56    3200        3887.33   9295.57   58.95   55.45   62.92    0.59
✔ Saved pipeline to output directory
utput/model-last
```

Figure 1. Training Result

Figure 1 shows the training pipeline of the model. Upon completion of the training, at 28 epoch and 1600 optimization steps, the model achieves a 62% accuracy score.

B. Testing Result



MATTHEW ELIOT

Summary

Senior Web Developer specializing in front end development. Experienced with all stages of the development cycle for dynamic web projects. Well-versed in numerous programming languages including HTML5, PHP OOP, JavaScript, CSS, MySQL. Strong background in project management and customer relations.

Education

Bachelor of Science: **Computer Information Systems** - 2014
Columbia University, NY

Skill Highlights

- Project management
- Strong decision maker
- Complex problem solver
- Creative Design
- Innovative
- Service-Focused

Experience

Web Developer - 09/2015 to 05/2019
Luna Web Design, New York

- Cooperate with designers to create clean interfaces and simple, intuitive interactions and experiences.
- Develop project concepts and maintain optimal workflow.
- Work with senior developer to manage large, complex design projects for corporate clients.
- Complete detailed programming and development tasks for front end public and internal websites as well as challenging back-end server code.
- Carry out quality assurance tests to discover errors and optimize usability.

Certifications

PHP Framework (certificate): **Zend, Codeigniter, Symfony**.
Programming Languages: **JavaScript, HTML5, PHP OOP, CSS, SQL, MySQL**.

Figure 2.1. Resume Test 1

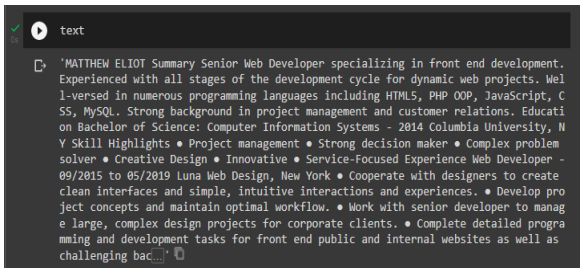


Figure 2.2. Resume Test 1 Extraction

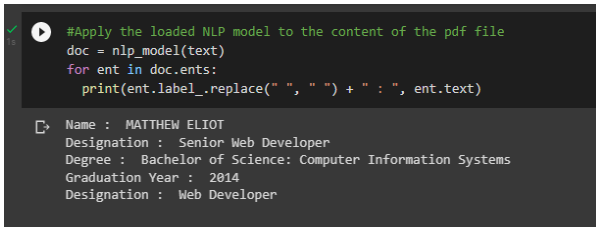
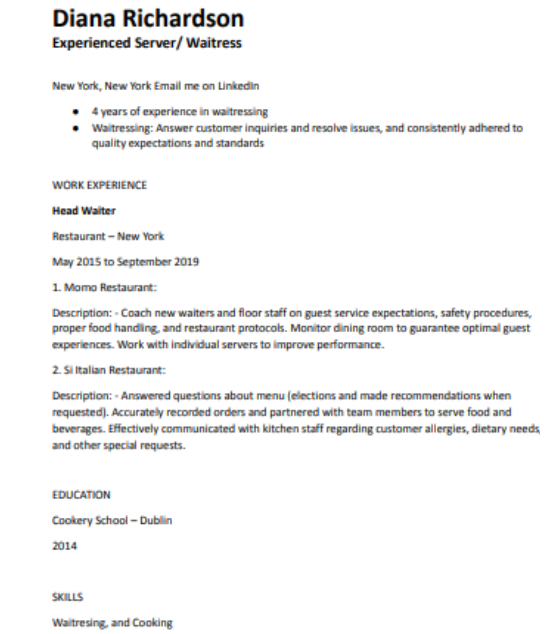


Figure 2.3. Resume Test 1 Result



Diana Richardson

Experienced Server/ Waitress

New York, New York Email me on LinkedIn

- 4 years of experience in waitressing
- Waitressing: Answer customer inquiries and resolve issues, and consistently adhered to quality expectations and standards

WORK EXPERIENCE

Head Waiter

Restaurant – New York

May 2015 to September 2019

1. Momo Restaurant:

Description: - Coach new waiters and floor staff on guest service expectations, safety procedures, proper food handling, and restaurant protocols. Monitor dining room to guarantee optimal guest experiences. Work with individual servers to improve performance.

2. Si Italian Restaurant:

Description: - Answered questions about menu (elections and made recommendations when requested). Accurately recorded orders and partnered with team members to serve food and beverages. Effectively communicated with kitchen staff regarding customer allergies, dietary needs, and other special requests.

EDUCATION

Cookery School – Dublin

2014

SKILLS

Waitresing, and Cooking

Figure 3.1. Resume Test 2

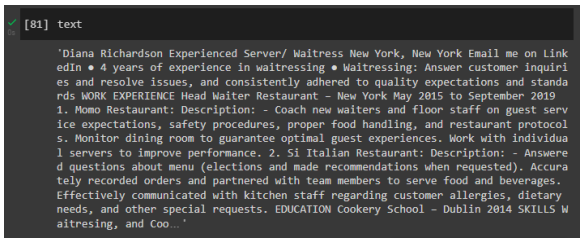


Figure 3.2. Resume Test 2 Extraction

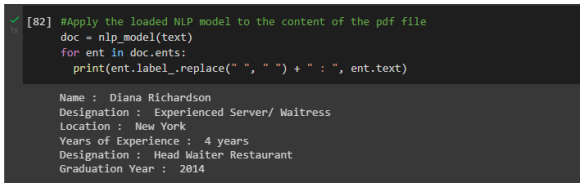


Figure 3.3. Resume Test 2 Result

The sample resumes that were used to evaluate the performance of the system are shown in Figure 2.1 and 3.1. On the other hand, Figure 2.2 and 3.2 illustrates the extracted information after testing the sample resumes to the model. Figure 2.3 and 3.3 shows the information in their respective categories.

This study has determined upon testing that the model's accuracy score is not the sole factor in the extraction and categorization of the information. The performance of the system also depends on the relevant and irrelevant count of words extracted from the resume. Thus, despite the relatively low accuracy score obtained after training the model, the extraction and categorization of the information from the sample resumes are correct and accurate, as shown in Figure 2.3 and Figure 3.3.

V. CONCLUSION & RECOMMENDATION

In this study, the researchers were able to successfully extract and summarize the important keys from a resume using NLP. The researchers advise future researchers to use a larger dataset with a higher quantity. Furthermore, because the researchers were unable to do so due to time constraints, the system should have a user interface and the display of the output could be improved further by having an output that is converted into a document file. Finally, future researchers could improve this study by focusing on much more complicated resumes.

VI. REFERENCES

[1] N.A. (2021). Why Is a Resume Important? (Types and Why You Need One). Indeed Editorial Team. https://www.indeed.com/career-advice/resumes-cover-letters/why-is-a-resume-important

[2] Selig, Jay (2022). The Role of Natural Language Processing (NLP) Algorithms. https://www.expert.ai/blog/natural-language-processing-algorithms/

[3]Smith, J. (n.d.). *7 Things You Probably Didn't Know About Your Job Search*. Forbes. Retrieved January 30, 2023, from https://www.forbes.com/sites/jacquelynsmith/2013/04/17/7-things-you-probably-didnt-know-about-your-job-search/?sh=4d27035e3811

[4] Bhushan Kinge, Shrinivas Mandhare, Pranali Chavan, & S. M. Chaware. (2022). Resume Screening using Machine Learning and NLP: A proposed system. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 253–258. https://doi.org/10.32628/cseit228240

[5] Gogulamudi, Vijay & Yadav, Arvind & Vishnupriya, B. & Lahari, M. & Smriti, J. & Reddy, D.. (2021). Text Summarizing Using NLP. 10.3233/APC210179.

[6] Sadiq, Sayed, et. al. (2016) Intelligent Hiring with Resume Parser and Ranking using Natural Language Processing and Machine Learning. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 4, p. 7437 – 7440.

[7] M. Gupta, "A Review of Named Entity Recognition (NER) Using Automatic Summarization of Resumes," *Medium*, Jul. 09, 2018. https://towardsdatascience.com/a-review-of-named-entity-recognition-ner-using-automatic-summarization-of-resumes-5248a75de175 (accessed Jan. 30, 2023).

[8] N. G. O and Hashwanth S, “Named Entity Recognition based Resume Parser and Summarizer,” *ResearchGate*, Mar. 31, 2022. https://www.researchgate.net/publication/359694475_Named_Entity_Recognition_based_Resume_Parser_and_Summarizer (accessed Jan. 30, 2023).

[9] S. Mali, “NAMED ENTITY RECOGNITION (NER) USING AUTOMATIC SUMMARIZATION OF RESUMES,” Peer-Reviewed, Open Access. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/issue_6_june_2022/25876/final/fin_irjmets1655128099.pdf

[10] Liu, Yinhan, et. Al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692