Resume Content Scoring and Improvement Suggestions Using NLP and Rule-based Techniques

8th International Conference on Information Technology Research 2023

Resume Content Scoring and Improvement Suggestions Using NLP and Rule-based Techniques

- Mr. R.L. Weerasinghe, Undergraduate, Faculty of Information Technology, University of Moratuwa
- Mr. N.N. Perera, Undergraduate, Faculty of Information Technology, University of Moratuwa
- Mr. S.P. Warusawithana, Undergraduate, Faculty of Information Technology, University of Moratuwa
- Ms. T.M. Hindakaraldeniya, Undergraduate, Faculty of Information Technology, University of Moratuwa
- Dr. (Ms.) G.U. Ganegoda, Senior Lecturer, Faculty of Information Technology, University of Moratuwa

Overview

- Introduction
- Contents of a Resume
- Data Source
- Dataset Preparation
- Analysis and Design
- Approach One for Content Scoring
- NER for Feature Extraction
- Approach Two for Content Scoring
- Missing Fields and Content Improvement Suggestions
- Conclusions

Introduction

- A strong resume increases the chance of getting selected for best job opportunities.
- Preparing an impactful resume requires paying attention to even the most intricate details as it would most probably be the point where the recruiting team would get the first impression of the applicant.
- When an individual is working on the same document for a long time it is typical that he/she could miss these details, so it requires a third party to review and spot them.
- Undergraduates have been unable to prepare a strong resume to match their skills and experience mainly due to the lack of the expert knowledge.
- Most of the existing solutions focus on how to shortlist candidates for a job position by identifying the talents of the candidate mentioned in the resume.
- It is important for candidates to have a proper tool to get their resumes analyzed and get feedback on content quality and missing contents. The primary objective of this study is to develop a tool for that purpose.

Contents of a Resume

- According to the studies, 26.1% of recruiters spend only 30-60 seconds and around 30% of the recruiters spend 1-2 minutes while 27.5% spend 2-3 minutes. Hence, it is important to mention content in a way to gain attention in such a short time
- As per the study, it is identified that the followings are the most important sections that should be included in a resume.
  - Personal information
  - Personal opening, job objective, career objective, and summary of qualifications
  - Education
  - Work experience
  - References
  - Scholarships, awards, honors
  - Hobbies, interests, and extracurricular activities
  - Willingness to relocate and travel

Data Source

- The resume list and their details were obtained from the system administrator of the industrial training platform of the Faculty of Information Technology, University of Moratuwa, Sri Lanka.

Analysis and Design

- When it comes to the development of this model, two major approaches were considered
  - Approach 01 - section-wise extracted resume content as a whole and the section-wise score were fed into the model for future predictions
  - Approach 02 - fed a dataset with section-wise extracted specific features and section-wise score to the model to predict the scores for unseen resumes

Approach One for Content Scoring

- This approach analyzes and provides scores for major sections based on the overall content of each section.
- Data Preprocessing
  - The dataset was cleaned by using basic NLP techniques and further by removing the section title, and additional spaces contained in the dataset
  - a BERT transformer was utilized, hence, commonly used preprocessing steps like stemming and lemmatization were not applied to the dataset.
- Feature Extraction
  - When it comes to giving a score for text content, having contextual meaning could increase the accuracy of the score.
  - Through the research, it was identified that Bidirectional Encoder Representations from Transformers (BERT) is more suitable for the implementation of this model.
  - Therefore, at the feature extraction step, BertTokenizer was used to produce word embeddings. In this model the 'bert-base-uncased' model is used as the base model.

Approach One for Content Scoring (Continue)

- Regression Model for Section Content Scoring
  - For content scoring regression model implemented using BertRegressionModel with an additional sequential layer
  - Dropout layer was added to prevent overfitting the model for training dataset.
  - Then a regression layer was added to predict the score by identifying the relationship between features and scores in the training dataset.
  - Then the model was trained using 40 epochs.
  - Mean Squared Error Loss function (MSELoss) was used as the loss function which measures the mean squared error between predicted score and actual score.

NER for Feature Extraction

- A Named Entity Recognizer (NER) was created in order to identify and extract particular features from the resume content.
- The resume information was labeled using “NER Annotator for spaCy” with predefined labels
- To train the Named Entity Recognizer (NER) model, the popular natural language processing library, spaCy, was utilized.
- The final model scored 0.99 as the score. This NER model could identify the section specified features successfully.

Approach Two for Content Scoring

- This approach involved developing individual models for each section to predict scores for the section content based on the section specified features
- Section specific feature extraction using NER
  - specific features considered for each section
  - Ex:

Section	Feature	Value Type
Profile Section	PROFILE DESCRIBING ADJ	No of person describing adjectives present in the section
	Content length	Length of the profile section
	DESIGNATION	Whether the job position that the candidate going to apply is included in the section or not

Approach Two for Content Scoring (Continue)

- Regression Models for section content scoring
  - several regression models were implemented, namely Linear Regression, Decision Tree Regression, Support Vector Regression (SVR), and Random Forest Regression. The objective was to identify the most suitable model for each section.

Section	MAE				
	Approach 01	Approach 02			
	BERT Regression	Linear Regression	Decision Tree	SVR	Random Forest
Profile	0.0667	1.7209	1.5581	1.5581	1.4186
Education	0.7333	1.8837	0.7442	1.0000	0.6279
Projects	0.5333	1.5349	0.9767	1.1860	1.1163
Technical Skills	1.2667	1.7714	2.2286	1.8286	1.7714

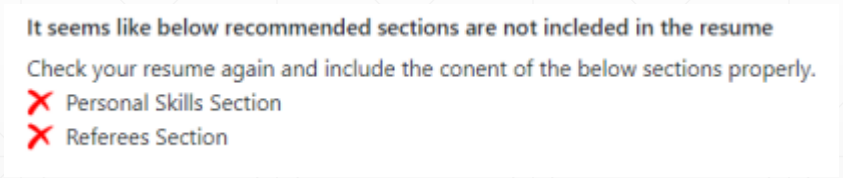
Approach One vs Approach Two

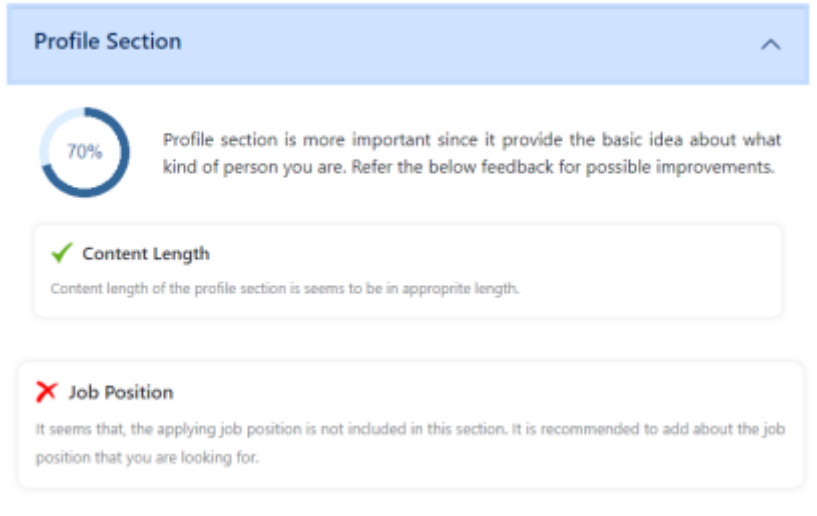
Based on the MAEs, for the profile, projects and technical skills sections, the BERT regression model developed in the approach 01 is more suitable where it considers entire content of the section and the context of the content when predicting the score for the section.

For the education section, random forest regression model which was developed in the approach 02 where it considers the section specified information is more suitable

Missing Fields and Content Improvement Suggestions

- This was implemented using a rule-based approach based on the features extracted from the NER model.
- As the first step, system provide list of sections as presented and unpresented sections (which are ideally should be included in a resume) on the given resume by checking the extracted contents using rule based approach.
- Then the system check the each section based on section specific features mentioned in the table 1 and provide feedback on each section using rule based approach to improve the contents of each section





Conclusions

- When analyzing the content of the resume, it is important to identify what are the important sections and details that should be included in the resume.
- When it comes to the scoring model, when considering the overall contents of a sections, for the feature extraction phase vectorization methods like CounterVectorizer in not much suitable.
- It is identified that BERT (Bidirectional Encoder Representations from Transformers) is more suitable for the scoring model since it also takes the context of the content into consideration.
- It is also identified that for sections like profile and projects of the resume to gain more accurate score it is needed to consider the whole content of the section where sections like education provide more accurate results when considering only the section-specified features.

Thank You