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

R.L. Weerasinghe
Faculty of Information Technology
University of Moratuwa, Sri Lanka
reshakalakshan@gmail.com

N.N. Perera
Faculty of Information Technology
University of Moratuwa, Sri Lanka
nimnaperera98@gmail.com

S.P. Warusawithana
Faculty of Information Technology
University of Moratuwa, Sri Lanka
supunawa@gmail.com

T.M. Hindakaraldeniya
Faculty of Information Technology
University of Moratuwa, Sri Lanka
tharumadurangi97@gmail.com

G. U. Ganegoda
Faculty of Information Technology
University of Moratuwa, Sri Lanka
upekshag@uom.lk

***Abstract* — Having a proper resume is very important for undergraduates or fresh graduates to find their dream job. But most of them find it difficult to prepare their resume properly by themselves. It often needs a third party to review the resume to identify missing parts and content improvements of the resume because most of the time candidates make some mistakes. When it comes to resume review systems, most of the systems are based on the recruiter perspective which does not provide any insights for the candidate to improve their resumes. Hence, it is helpful if a proper resume content reviewer is there for candidates to analyze their resumes. This study is focused on developing a model to resume content scoring and suggest missing content based on NLP and rule-based techniques. Two separate approaches were developed and tested for the proposed system and then the comparison of those approaches were carried out through this study.**

Keywords* — *Resume Content Scoring, Missing Content Suggestion, NLP, Rule-based Techniques

I. INTRODUCTION

With the recent developments in the education system, there are many undergraduates in the IT industry who are looking for opportunities in the field of IT. However, due to limited openings in IT companies, they can only accommodate a few opportunities for undergraduates, resulting in intense competition among undergraduates. When it comes to winning this competition, the resume plays a huge role. A strong resume increases the chance of landing the best opportunity. 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 preparing a resume, a number of factors need to be considered such as its structure, content, design, grammar, and spelling [1]. An undergraduate can easily be confused about what content to include and how to properly structure them leading them to either miss out on important details or include unnecessary details. The structure should be such that it highlights the important points and is easy to read through. Having proper grammar and no spelling mistakes is another important factor when it comes to a strong resume. 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.

Even though a strong resume is a powerful tool for a candidate [2] undergraduates must go through several steps to create a strong resume. And the last and hardest step is reviewing the resume and it requires expert knowledge. When looking at the past, it can be seen that a lot of undergraduates have been unable to prepare a strong resume to match their skills and experience mainly due to the lack of this expert knowledge. Yet there is no proper solution on the internet for getting feedback on the resumes prepared by undergraduates [3].

There are several research which have been done for analyzing the resume content. However, most of them focus on how to shortlist candidates for a job position by identifying the talents of the candidate mentioned in the resume. Those approaches didn't focus on scoring the resume content based on quality and content. The primary focus of those approaches was to extract skills and other prominent information and rank the candidate based on that data.

In the study by Hewage et al. [4] a method has been proposed to create resumes using Gensim LDA unsupervised machine learning model according to the Hoffman method. This method only gives an overall rating for the resume by only considering the relevant skills and particular job. It does not rate the content of each section.

In the study by Zimmermann [5], scores are given for only the extracted content of the education, skills, and work experience sections based on the predefined score criteria by considering the job description. This approach does not consider the quality of the overall content of the section, only paying attention to the contents they are looking for. For candidates to create a proper resume, they need to have quality content in each section.

The study by Shestakova [6] proposed a method to analyze the contents of the resume and recommend job vacancies by considering the similarities between the job description and resume content. BERT and DistilBERT models were applied when developing this approach. To measure the similarity, the method has utilized the cosine similarity approach. Even though this approach analyzes the contents of the sections, it doesn't provide any feedback about the quality of the contents of each section.

When it comes to the undergraduate perspective, it is important for them 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.

The rest of the paper is laid out as follows: The first section includes about the contents of a resume. Then the methodology of the proposed system in the next section. Finally, conclusion is included in the last section of the paper.

979-8-3503-5950-3/23/$31.00 ©2023 IEEE

II. CONTENTS OF A RESUME

In relation to the content of a resume, there are numerous studies which were carried out throughout the past few decades. Contents of a resume play a major role when getting recruited for a job position. The First impression about the job applicant is usually made through the content of the resume [7]. Hence it is important to have good and quality content on the resume. In the study by Burns et al [7], it was found that there exists a correlation between the content included in a candidate's resume with his/her personality.

Schramm et al. [8] carried out research about what constitutes the ideal resume by collecting information through a questionnaire from recruiters. This research pointed out the importance of the content, format and appearance of a resume. When it comes to studying resumes by a recruiter, as shown in the results, 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. According to the study, 90.8% of recruiters agreed that the permanent address should appear on the resume. When it comes to the education details, almost all the recruiters agreed that the college where the candidate graduated should be mentioned in the resume. Also, more than 82% of recruiters mentioned it is better to add the other colleges attended to the resume. Almost every recruiter mentioned that work experience should be included in the resume while 94% of the recruiters indicated that the date of employment also should be there. In addition to that, 74% of the recruiters preferred to have an explanation of the work experience as well. Moreover, more than 85% of the respondents concurred that extracurricular activities as well as achievements ought to be listed in the resume. According to the study, when considering the length of the resume, a maximum of two pages was favored by two thirds of the recruiters. Between 70% and 80% of the recruiters concurred that a disinterest in a prospect will result from poor arrangement and a lengthy resume. More than 80% of those surveyed gave the order in which the things appeared on a resume some level of significance. When it comes to the grammar and other mistakes, 95% of respondents believed that more than one spelling mistake or poor grammar will make some recruiters less interested in a candidate. More than 80% of the recruiters said that having multiple typing errors on a resume would make it more difficult for the applicant to have an interview. Through this study it has identified the importance of quality content, format, and order of a resume as mentioned above.

In the study by Risavy [9] also mentioned the importance of the resume content. According to that study, the following information 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

Additionally, this study mentioned what are the most important details that should be included in the above sections. According to the research applicant's name, phone number and address are mandatory information that should be included as personal information. Moreover, it has been identified that early and more contemporary resume research both concur, that formal educational qualifications, such as degree or the designation and the major, minor, and if applicable, the anticipated graduation date should be listed on resumes. Moreover, according to this study, candidates should include information about their former jobs, including the dates they worked there, the position they had, and whether or not it was a full or part-time one. Apart from that information, it also mentioned about the length of the resume and the order of the contents. When considering the length of the resume, this study also mentioned that three-page resumes should be avoided in favor of one-to-two-page resumes. Moreover, this study mentioned that the content should better be in the order of personal information, education, work experience, and extracurricular activities. Additionally, it mentioned that details such as height, weight, race, religion, birth date, marital state, number of dependents, physical/health status, and social security number should not be included in the resumes.

According to both early and recent research, it is important to have good content, ordered perfectly within one or two pages when preparing the resume. Because, a resume is the first thing that creates an image of the applicant on the recruiters head and recruiters only spend very little time to go through a resume. Therefore, it is important to gain the attention of the recruiter for the resume within a few seconds.

III. METHODOLOGY

The target users of the proposed system have lesser knowledge on how to properly write the content of the resume as this would be most likely their first experience in building a professional resume. Also, they could sometimes miss adding important sections to the resume and make grammar mistakes when preparing the resume.

To tackle this issue, a model was implemented to give a score for each section of content based on the quality of the content. When considering the process of scoring the content quality, it cannot be addressed as a classification problem since it may get continuous values. Hence, a regression model was used to give a score for the content quality which would be more accurate. Further, a rule-based approach is used to suggest missing sections.

A. Data Source

Several different data collection methods had to be adopted in order to collect data to suit the requirements of different modules. To successfully train the proposed models, a large number of resumes of previously selected candidates were required. 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.

B. Dataset Preparation

After obtaining resumes, the contents of each section of those resumes were scored by the HR professionals from the

industry. And then, those scored dataset was utilized for the training and validation of the proposed model.

Moreover, label studio software was utilized to annotate the resume contents according to the relevant sections and then those data were extracted and prepared an excel sheet along with the scores obtained.

C. Analysis and Design

When it comes to the development of this model, two major approaches were considered. In the first approach, section-wise extracted resume content as a whole and the section-wise score were fed into the model for future predictions, while in the second approach, it was fed a dataset with section-wise extracted specific features and section-wise score to the model to predict the scores for unseen resumes. The design plan for the proposed solution is shown in Figure 1.

D. Approach One for Content Scoring

This approach analyzes and provides scores for major sections based on the overall content of each section.

1) Data Preprocessing: When implementing the model, as a first step it is required to preprocess the created dataset for further usage. The dataset was cleaned by using basic NLP techniques and further by removing the section title, and additional spaces contained in the dataset. In the current implementation, a BERT transformer was utilized, hence, commonly used preprocessing steps like stemming and lemmatization were not applied to the dataset. In addition to that, the score was converted to an integer.

2) Feature Extraction: As a first approach, CountVectorizer was used which transforms text content into vectors by considering the frequency of each word occurring in the entire dataset. Since it considers the frequency of the words, it couldn't capture the contextual meaning of the words. When it comes to giving a score for text content, having contextual meaning could increase the accuracy of the score. For example, when it comes to the profile section, someone can introduce themself in a more unique way which is eligible to have a higher score. But when only considering the word frequencies, that content could get a low score due to the fact that the words used in that content have less frequencies than the words used in the contents which got more score. Therefore, it is identified that the contextual meaning also plays a major part when giving a score for text content.

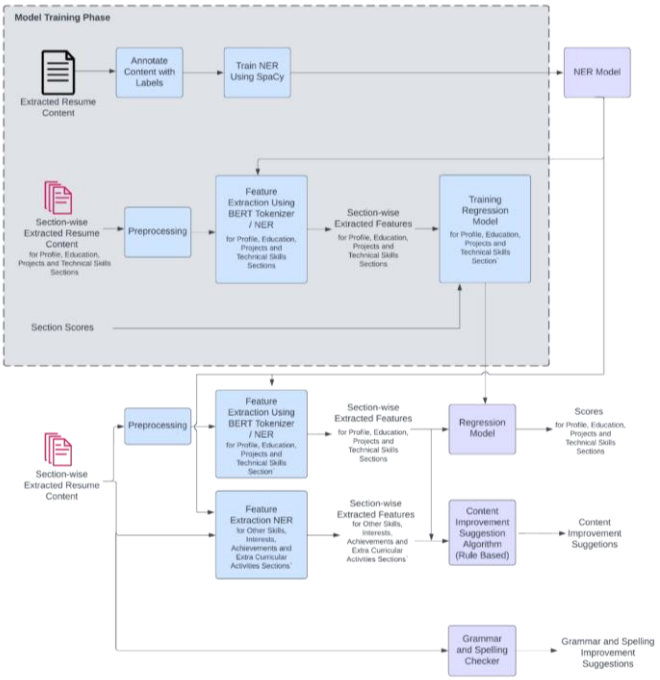

Figure 1: *High-level Architecture of the proposed model*

Through the research, it was identified that Bidirectional Encoder Representations from Transformers (BERT) [10] is more suitable for the implementation of this model. It is a pre-trained neural model which follows transfer learning mechanism and produces word embeddings which could be used as features. BERT uses Transformer, an attention mechanism that recognizes the relationships between words in a text based on contextual meaning. The model is implemented based on BERT regression. 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.

3) Regression Model for Section Content Scoring: For content scoring regression model implemented using BertRegressionModel with an aditional sequential layer. In this model a 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 t*he 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.

E. NER for Feature Extraction

When analyzing resume contents, considering the specific information in the resume is very important. A Named Entity Recognizer (NER) [11] was created in order to identify and extract particular features from the resume content. The NER's goal is to locate and extract pertinent information for each section of the resume.

The resume information was labeled using "NER Annotator for spaCy" with predefined labels that matched the various features specified to each section in resumes in order to train the NER. After creating the labeled dataset by manually annotating the resumes, the next step was to train the NER to extract the features using unseen resume contents. To train the Named Entity Recognizer (NER) model, the popular natural language processing library, spaCy, was utilized. spaCy provides a convenient and efficient framework for training custom NER models. Prepared dataset should be converted into spaCy format prior to training the model. A DocBin object must be built in order to store our data before converting it to spaCy format. After adding the example and the entity label to the DocBin object, then iterate through the data and save the object to .spacy file. The final model scored 0.99 as the score. This NER model could identify the section specified features successfully.

Prepared dataset should be converted into spaCy format prior to training the model. A DocBin object must be built in order to store our data before converting it to spaCy format. After adding the example and the entity label to the DocBin object, then iterate through the data and save the object to .spacy file.

Once the pipeline had been trained, the best model was saved in the output directory. The final model scored 0.99 as the score. This NER model could identify the section specified features successfully.

F. 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. The initial phase of this approach focused on feature extraction of each section, utilizing the previously developed NER model.

1) Section Specified Feature Extration Using NER: Table 1 outlines the specific features considered for each section. Similar code segments were developed to extract features from each section by considering the features applying to each section.

2) Regression Models for Section Content Scoring: Following the feature extraction process, the subsequent step involved training regression models for each section to predict scores for the corresponding section contents. To accomplish this, 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.

Following the training phase, the models are evaluated using several metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared score (R2 score).

G. Missing Fields and Content Improvement Suggestions

In addition to providing scores for the profile, education, projects and technical skills sections, this solution will provide some suggestions to improve the resume by considering the missing contents or sections and possible content improvements. 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 setion specific features mentioned in the table 1 and provide feedback on each section using rule based approach to inprove the contents of each section.

Table 1: Specific features considered for each section

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
Education Section	DGREE	Whether the degrees that the candidate currently reading is included in the section or not
	GPA	Value of the GPA if included in the section
	GRADUATION YEAR	Whether the graduation year of the candidate is included in the section or not
	SCHOOL	Whether the school of the candidate is included in the section or not
	AL STREAM	Whether the A/L Stream of the candidate is included in the section or not
Projects Section	PROGRAMMING LANGUAGES	No of programming languages present in the section
	DATABASE	No of database technologies present in the section
	WEB DEVELOPMENT	No of web development technologies present in the section
	BACKEND SERVICES	No of backend services present in the section
	MOBILE APP DEVELOPMENT	No of mobile app development technologies present in the section
	OTHER TECHNOLOGIES	No of any other technologies present in the section
	PROJECT KEYWORDS	No of project key words present in the section
Technical Skills Section	PROGRAMMING LANGUAGES	No of programming languages present in the section
	DATABASE	No of database technologies present in the section
	WEB DEVELOPMENT	No of web development technologies present in the section
	BACKEND SERVICES	No of backend services present in the section
	MOBILE APP DEVELOPMENT	No of mobile app development technologies present in the section
	PROJECT MANAGEMENT TOOL	No of project management tools present in the section
	IDE	No of IDEs present in the section
	VERSION CONTROLLING	No of version controlling technologies present in the section
	OTHER TECHNOLOGIES	No of any other technologies present in the section

IV. Discussion

This section includes evaluation and comparison of the proposed approach one and two.

When considering the scoring the section contents, total of five models were developed for each section: one model from Approach 01 and four models from Approach 02. The evaluation of these models will help determine the best-performing model for each section. The main difference between the two approaches is that Approach 01 takes into account the entire content of the section, while Approach 02 focuses on specific features extracted through the NER model. Both approaches utilize regression models, with Approach 01 using the BERT regression model, and models in Approach 02 employing linear regression, decision tree regression, SVR, and random forest regression.

Since all five scoring models predict scores between 0 and 10, it is more appropriate to use Mean Absolute Error (MAE) rather than R-squared (R2) score for comparing models. Here's why:

- Interpretability: MAE provides a direct interpretation of the average prediction error in the original units of the target variable. In this case, the scores range from 0 to 10, and the MAE would represent the average absolute difference between the predicted and actual scores. This gives you a clear understanding of the average prediction error in the context of the scoring scale.
- Magnitude of Errors: MAE is sensitive to the magnitude of errors, as it calculates the average absolute difference between predictions and actual values. It will directly reflect the size of the errors in the score predictions. This is important when the absolute difference between predicted and actual scores is more critical than the direction of errors.
- R2 score interpretation: R2 score measures the proportion of variance in the target variable that is explained by the model. However, the interpretation of R2 score is less intuitive when the target variable has a limited range, like scores between 0 and 10. R2 score may not provide a meaningful representation of the model's performance in this specific context.

As explained, for a scoring model where the predicted scores range from 0 to 10, using MAE is more appropriate for comparing models. It provides a straightforward interpretation of the average prediction error in the score scale and is more suitable when the absolute difference between predicted and actual scores is of primary importance. Hence MAE was utilized to evaluate the models developed through approach 01 and approach 02. The table 2 shows the MAEs for each model trained for each section.

Table 2: MAEs for each model trained 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

Based on the MAEs shown in the above table, 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.

Table 3 shows the scores predicted for the given sample content for each section from models developed through both approach 01 and approach 02.

Table 3: Scores predicted for the given sample content for each section.

Section	Content	Score					
		Actual	Approach 01	Approach 02			
			BERT Regression	Linear Regression	Decision Tree	SVR	Random Forest
Profile	Ability to handle highly stressed situations and unexpected errors in an efficient manner	02	02	02	02	02	02
	Dedicated third year undergraduate with strong interpersonal skills and extensive knowledge in the field of Information Technology, seeking for an internship in an organization that indulges personal growth while allowing me to utilize my knowledge and skills.	07	08	07	07	07	07
Education	BSc. (Hons) in Information Technology — University of Moratuwa Reading: second year - CGPA: 3.12 (Level 1) G.C.E. Advanced Level 2017 Z-Score: 1.5013 Central College, Anuradhapura, Sri Lanka	06	06	07	06	04	06
	B.SC. (HONS.) IN INFORMATION TECHNOLOGY UNIVERSITY OF MORATUWA Level 1 Semester 1 GPA 3.78 Level 1 Semester 2 GPA 3.88 Level 2 Semester 1 GPA 3.56 Overall GPA 3.72 HOLY FAMILY CONVENT, COLOMBO 04 G.CE (A/L - 2013) - PHYSICAL SCIENCE STREAM Results - Physics A, Combined Mathematics B, Chemistry B GCE (O/L - 2009) Results - 6As 2Bs 1C PROFESSIONAL QUALIFICATIONS Successfully completed a Certificate course in Computer Science at National Institute of Business Management (NIBM), 2010	08	08	07	08	08	08

Projects	class manager system for addressing a website and a mobile application to tuition classes for ease of their activities	02	01	03	02	02	02
	Remote mariner an under water camera that can be controlled by a remote controller and get the visual feedbacks. technologies: pic programming, wireless networking weather - now mentored by virtusa polaris(pvt)ltd. a crowdsourcing web application and android application which can be used to know real time weather updates and also it can be updated according to current weather satus and get the personal ranks of reliable updates. projec link: matrix.projects.mrt.ac.ik:3000 technologies: angularjs, nodejs, html5, mongodb google maps api, android, mysql	07	07	06	07	06	06
Technical Skills	Web Development HTML, CSS, JavaScript, Bootstrap, jQuery #NAME? Version Control Systems Git	05	06	04	05	03	04
	Programming Languages - C, Java Databases - MySQL, Oracle Web designing - HTML, JSP, JavaScript, php, Bootstrap, Servlets Multimedia - Sketchup IDE - Netbeans, Eclipse, Atmel Studio Version Control System - GitHub Other - Rational Rose, Star UML	08	09	06	06	08	08

V. CONCLUSION

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. Because it only considers the frequencies of the words presented. It is also important to consider the context of the information to provide more accurate results. 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.

This study encountered limitations due to the scarcity of available data resources, which posed challenges in obtaining a sufficiently large and precise dataset. The primary limitation identified in this research pertains to the practical difficulty in acquiring highly accurate and specific data. The scores assigned to different sections of the resumes tended to exhibit a bias towards scores higher than 5. To mitigate this bias and improve the dataset's representativeness, additional data was incorporated with the assistance of recruiters. It is important to acknowledge that the models employed in this study were trained using limited data resources. Consequently, the outcomes and performance of these models are contingent upon the dataset used for training.

As for further research studies based on this research, for the scoring model, another approach where it analyzes the context of the section-specified features could be tried. Also, an approach with the combination of both overall content and section-specified features could be studied to find whether those approaches would have any impact on the accuracy of the results. Moreover, the performance of this module could be enhanced by training the models with larger datasets.

ACKNOWLEDGMENT

This study is supported by the Faculty of Information Technology of the University of Moratuwa under the supervision of the Department of Information Technology and the Department of Interdisciplinary Studies.

REFERENCES

[1] "7 simple but effective ways to make your CV stand out | Top Universities." https://www.topuniversities.com/blog/7-simple-effective-ways-make-your-cv-stand-out (accessed Sep. 06, 2023).

[2] J. Rout, S. Bagade, P. Yede, and N. Patil, "Personality Evaluation and CV Analysis using Machine Learning Algorithm," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 5, pp. 1852–1857, May 2019, doi: 10.26438/ijcse/v7i5.18521857.

[3] P. G. Roos, "Development and evaluation of a competence-based curriculum vitae-writing programme for new graduates," North-West University, 2018.

[4] H. Hewage, K. U. Hettiarachchi, K. Jayarathna, K. P. C. Hasintha, A. N. Senarathne, and J. Wijekoon, "Smart human resource management system to maximize productivity," in *2020 International Computer Symposium (ICS)*, 2020, pp. 479–484.

[5] T. Zimmermann, L. Kotschenreuther, and K. Schmidt, "Data-driven HR - Résumé Analysis Based on Natural Language Processing and Machine Learning," *CoRR*, vol. abs/1606.05611, 2016, [Online]. Available: http://arxiv.org/abs/1606.05611

[6] A. Shestakova and A. Corradini, "Exploring the Use of Machine Learning for Resume Recommendations," in *International Conference on Speech and Computer*, 2022, pp. 626–640.

[7] N. H. Anderson and A. A. Barrios, "Primacy effects in personality impression formation.," *The Journal of Abnormal and Social Psychology*, vol. 63, no. 2, p. 346, 1961.

[8] R. M. Schramm and R. Neil Dortch, "An analysis of effective resume content, format, and appearance based on college recruiter perceptions," *The Bulletin of the Association for Business Communication*, vol. 54, no. 3, pp. 18–23, 1991.

[9] S. D. Risavy and others, "The resume research literature: Where have we been and where should we go next," *J Educ Develop Psychol*, vol. 7, no. 1, pp. 169–187, 2017.

[10] S. Alaparthi and M. Mishra, "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey," *arXiv preprint arXiv:2007.01127*, 2020.

[11] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, "Named entity recognition approaches and their comparison for custom ner model," *Sci Technol Libr (New York, NY)*, vol. 39, no. 3, pp. 324–337, 2020.