Improving Attractiveness of a Job Advertisement

K. Alkin Gunbay, Ecem Ture, Olga Beckchanova
s288625@studenti.polito.it s288621@studenti.polito.it s275308@studenti.polito.it
Politecnico Di Torino
Turin, Italy

***Abstract*—The number of employees and employers is growing exponentially. Human Resource Departments of the companies are struggling to reach the best matching candidates for their open job position. Job advertisements are key to attracting more and best matching candidates. In this study, we aimed to create an algorithm that enhances the attractiveness of a Job Description by suggesting more attractive skillsets to the writer. The enhanced version of the Job Description will attract more and best matching candidates. We evaluated our approach in a recent Job Advertisements Dataset. We showed that it's possible to define a metric of attractiveness and generate a list of more attractive skills compared to the given input.**

I. Introduction

Recruiters and HR departments of companies receive so many applications every day and they spend their most of time examining applicants' Resumes. This situation may cause mistakes such as unnoticed qualified candidates. Also, a well-written job advertisement is important as a Resume because a well-written job ad increases the attention to the position and makes it easy to reach the best candidate. In addition, applicants prepare different types of Resumes and add unrelated skill definitions. We studied to find a solution for these issues in collaboration with the INDA (which is a company that studies on the machine learning solutions in Job Advertisement market).

We aim to improve the attractiveness of job advertisements to use for the job advertisement writing process and provide the best matching between recruiters and candidates. As a result, the correct writing of a job advertisement will attract the attention of potential and appropriate job candidates and influence them to apply for a proper job. The main issue to overcome is the terms of measurement of the attractiveness of the skill. To be able to suggest more attractive skills to the recruiters, data science methods should be applied. In such a way, more accurate attractive skills would be offered to recruiters. To implement our solution, INDA provided two datasets to implement the solution. The first dataset is recent job advertisement data published on online platforms. This data needs pre-processing and cleaning to obtain enhanced data. Each job advertisement contains information such as job advertisement title, company name, job description, seniority level, job function, location, industry, date, and employment type. Another dataset is a skills data sample of INDA'S private skills database.

The solution consists of several components as pre-processing and cleaning of job advertisement datasets. Also, we need to deal with missing values and provide a statistical overview for features. We used the skill extractor model, which is developed by INDA, to extract the skills from the description of job advertisements. In this project to develop our solution, the focus is set on finding attractive and similar words inside the job advertisements to enhance effectiveness by the internal text improvement. To achieve our aim, we isolated features and focused on the relation between the Number of Applications and Job Description to define the attractiveness of a skill. The finally attractiveness score can be used as a metric to suggest a better word to the writer.

II. Related Work

The job market is continuously changing according to globalization and technology.

During the last decade, several researchers have been made about several different aspects of recruitment to understand job market needs as job descriptions and skill identifications. The authors [1] suggest using the skill extractor to extract skill sets from Job Adds to analyze the current and future state of the research market which is a similar method with our project. In another study[2], researchers developed a model that can match the candidate with proper job advertisement according to filtering metrics such as posting times and geolocations of job ads. This idea supports our approach of using the duration time of job ads in the normalization process because it shows that duration can be related to the number of applications. Different models have been used to find the similarity between words such as a model based on *HowNet*[3], *CiLin* and *Word2vec*[4]. *HowNet* uses the layer calculation method, *CiLin* uses the distance between words and distance branches and *Word2vec* model calculates the semantic similarity between words. After our research, we decided to use the Spacy Similarity function which uses the *Word2vec* model to find similarity.

III. method

Finding the best matching employee for a job position is the optimal goal of employers. To achieve this goal, companies have to attract the best matching candidates to their Job Advertisements. A Job Advertisement can be written to attract the best-matching candidates for that position. However, there isn't a template to increase the attractiveness of a position. Because of that companies are not able to reach the best matching candidates. To increase the attractiveness of a Job Advertisement, there should be an assistant tool to help the HR responsible. This tool can be created by using Data Science methods. This way a Job Advertisement can always be in its most attractive state to reach the best-matching candidates.

To complete our tool to enhance Job Advertisements, we used the Data Science pipeline shown by the Figure 3. Our task has two sub-goals as creating a metric of attractiveness and similarity. To achieve these goals, we used a data set provided by INDA. They used a data-scrapping tool to generate this data-set from a popular Job Advertisement platform. This data-set is formed by 10 features as; Number of Applications, Industry type, Company Names, Employment Type, Job Title, Job Description, Job Functions, Language of the Advertisement, Seniority Level, Location. We also have the metadata (Creation date of the Advertisement, Scrapping date of the Advertisements) information of the data-set. In the first step of our implementation, we started by pre-processing and analyzing our data.

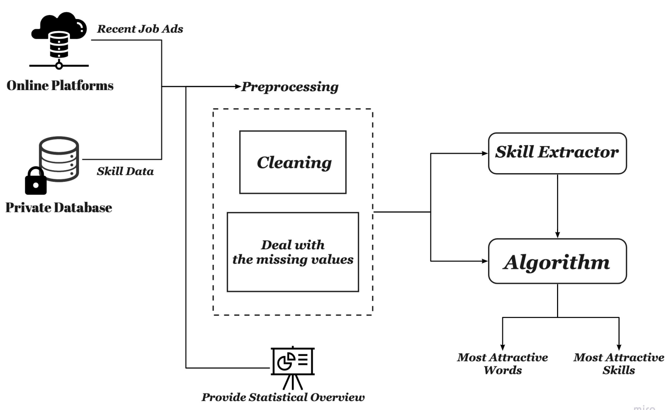

Fig. 1. A summary of our Data Science Pipeline

This data set contains 30.000 Job Advertisements. Figure 4 shows the number of empty cells in each feature. In our data analysis pipeline, we started by handling those empty cells. The Number of Applications category is the most relevant feature with the Attractiveness of the Job Advertisement. Because of that, we can not try to fill those cells. So, we are dropping every line with an empty Number of Applications. The Seniority Level has sub-categories as; Entry-Level, Mid-Senior Level, Associate, Director, Internship, and Executive. Although some of the Seniority Levels are empty, information to fill the missing Seniority Levels can be extracted from the Job Title. For example, if the job title is Senior-Developer, we can fill the blank Seniority level with "Mid-Senior Level". After filling those lines we are dropping every other row with an empty value. At the end of this process, we have 16.860 data points remaining which is ready for the statistical analysis.

When we investigated the location and language features, we observed that all the Job Advertisements are English. Also, they are in the United Kingdom. After that, we used the creation date and scrapping date to create duration information of the data. We observed that jobs advertisements are open for an average of 20 days. However, some advertisements can remain open for up to 80 days. Some of them got closed in 3 to 4 days. When we look at the categorical features (Seniority Level, Employment Type), we observed that 15.434 of all advertisements are looking for a Full-Time employee. Also, most of the applications (711.119) were received by full-time job advertisements. Around 60.000 Applications were received by the other employment types. On the other hand, the Seniority Level category has a more balanced distribution. There are 5000 Job Advertisements in the category of Mid-Senior Level. The number of advertisements for the Associate Seniority Level is almost 4000. After this analysis, we are ready to start working on the normalization of our dataset.

Since we need to create attractiveness of a Job Advertisement metric, we assumed that the Number of Applications received by each job Advertisement should be related to the Attractiveness of the Job Advertisement. However, this value is biased by the Company names, Employment Type, Job Title, Seniority Level, Duration of the Advertisement, Job Functions, and Job Descriptions. In our approach, we are trying to define the attractiveness of the Job Description itself. To achieve this, we need to remove the effect of other features on the Number of Applications. To achieve this, we developed a normalization method. We assumed that each categorical feature must have a coefficient that defines its effect on the Number of Applications. This coefficient is defined as the number of applications received by that Category divided by the Frequency of the advertisement. Figure **??** shows the calculation of the coefficients. The division operation allows us to remove the effect of the unbalanced distribution between categories. Finally, by using those arrays of coefficients we can normalize the Number of applications receives by each advertisement. This final normalization step will give us the relation between Job Description and the Attractiveness of the Textual Data. An example of this step shown by Figure **??**. After that, we used the skill extractor software to extract skill sets from the Job Description.

$$\textbf{Coefficient} = \left\{ \frac{\textbf{Number of Applications per Category}}{\textbf{Frequency of the Category}} \right\}$$

Fig. 2. Formula to calculate a coefficient for features

$$\textbf{Attractivity} = \textbf{Number of applications} * \frac{\textbf{1}}{\textbf{Duration}} * \frac{\textbf{1}}{\textbf{Coef}_{\textbf{1}}} * \ldots$$

Fig. 3. The Formula of the Attractiveness

To generate a skill-set from the Job Descriptions, we are using a deep learning model trained and developed by the INDA. Their model returns a list of skills generated from the given paragraph. After generating a skill-set for each Job Advertisement, we assigned an attractiveness score to those skills by using a cumulative averaging approach. Each skill is receiving a different score from the different advertisements and to calculate the final score for that skill we are averaging the total score for that skill. In this way, we can generate a skills dictionary that is formed by skills and their attractiveness. After that, we used the Word2Vector approach provided by the Spacy NLP library to measure the similarity between the given keyword and each skill located in the skills set. At this point, we can suggest a similar but more attractive skill to the final user to improve their content.

During our implementation, we used the Python programming language. We implemented our solution on the Google Collaborative platform. We used pandas and seaborn libraries to analyze and process our data. Finally, we used the Spacy NLP library to find similarities between the words.

IV. Experiment

In our study, we are using a dataset that is collected from the popular job advertising platform of LinkedIn. This dataset is scrapped by using Python libraries in logout mode. So the data is limited with LinkedIn's default features. This data set contains 10 main features as; Number of Applications, Industry type, Company Names, Employment Type, Job Title, Job Description, Job Functions, Language of the Advertisement, Seniority Level, Location. The Location feature has some sub-features as City, Country, Country code, County, GeoCoordinates, and Region. We also have other features such as source link, domain id, creation date, and posted date of the advertisement. The creation date defines the scrapping date of the data and posted date is the date that advertisement is published online. The dataset contains 30.000 entries. All of the advertisements are published in English and their location is in the United Kingdom. The Number of Applications feature is the most relevant feature to generate the attractiveness feature. Overall this dataset contains all the features that form a job description. Thanks to the variety of different features, we can observe the relationship between each feature and the Number of Applications.

Although the data set contains 30.000 entries, not all of them are useful for us. Figure 4 shows the number of empty cells in the dataset. Since the Number of Applications is our target value to the process, we can't work with NaN values are located in that feature. Because of that, we have to drop all rows that have NaN value in the Number of Applications feature. On the other hand, we can extract information from the Job Title or Job Description to fill NaN values of the Seniority Level. After dropping the rest of the values that we can't fill the remaining Number of Records is 16860.

In our study, we assumed that the number of applications defines the attractiveness of the advertisement. After that, we processed that feature, to calculate the attractiveness of the Job Description. Since there isn't any ground truth in our study, we evaluated the results by investigating them ourselves.

We considered two points as an outcome. First of all, since we are calculating the attractiveness of a Job Description, the Description with the highest attractiveness should have an extraordinary Description. When we take a look at that advertisement we observed some features that you can expect

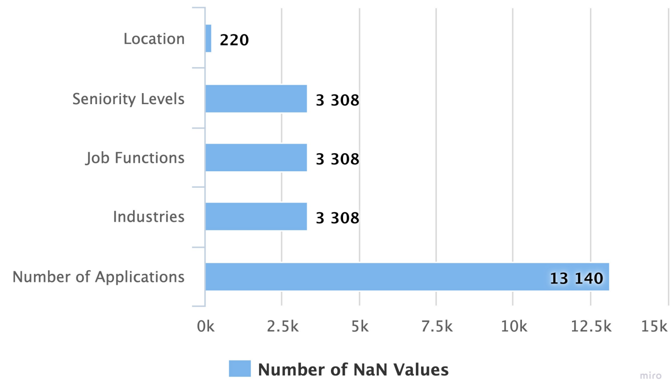

Fig. 4. Features with NaN values vs Number of NaN values

from a dream job. Here are some bullet points from the Advertisements with the highest score:

1) Possibility to work remotely
2) Work 2 months as part-timer with a paid holiday in a European country
3) 1 month vacation in a year
4) Regular cultural and social activities organised by a party committee

The second expected outcome is generating a more attractive skillset from the given skill by the user. In this experiment, we received the following results.

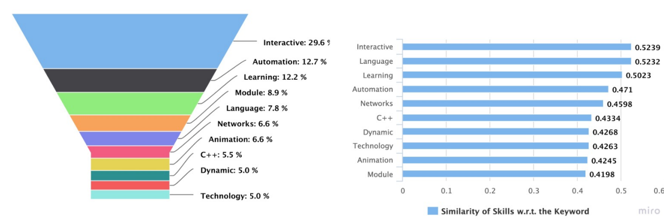

Fig. 5. Attractiveness and Similarity of the Suggested Skills

Figure 5 shows that the overall idea of suggesting more attractive skills to the user is working. However, newly generated skills might not be so relevant or meaningful. The reason why for those irrelevant or meaningless skills is the way we extracted them from the description of the advertisement. That method generates a list of single words. However, some skills are formed by two or three words such as "Project Management". Also, while assigning an attractiveness score, we can use more advanced techniques to improve the results such as assigning a coefficient to the skills. Otherwise score of a meaningless skill might be higher just because it's been used in attractive descriptions.

V. Conclusion

In this study, we aimed to develop a method to increase the attractiveness of a Job Advertisement by using Data Science methods. We manage to show that it's possible to define a metric of attractiveness by using a dataset created from Job

Advertisements. Also, we manage to suggest more attractive skills to the user, by using extracted skills from the Job Description and their assigned score of attractiveness. The content of a text can be improved to attract more relevant candidates by using our method. The same method can also be used to pick the most attractive descriptions from a corpus. The outcome of this study can be used to create an automated text enhancer. In this way, instead of generating just a set of skills, we can suggest a format to the user and improve the attractiveness of their paragraphs.

References

[1] Imane Khaouja, Ismail Kassou, and Mounir Ghogho. "A Survey on Skill Identification From Online Job Ads". In: *IEEE Access* 9 (2021), pp. 118134–118153. DOI: 10.1109/ACCESS.2021.3106120.

[2] Thomas Schmitt et al. "Language Modelling for Collaborative Filtering: Application to Job Applicant Matching". In: *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. 2017, pp. 1226–1233. DOI: 10.1109/ICTAI.2017.00186.

[3] Xuexia Gao and Tao Kuang. "Research of word similarity model based on HowNet". In: *Proceedings of 2011 International Conference on Electronic Mechanical Engineering and Information Technology*. Vol. 1. 2011, pp. 287–290. DOI: 10.1109/EMEIT.2011.6022926.

[4] Yushang Mao, Guixuan Zhang, and Shuwu Zhang. "Word Semantic Similarity Based on CiLin and Word2vec". In: *2020 International Conference on Culture-oriented Science Technology (ICCST)*. 2020, pp. 304–307. DOI: 10.1109/ICCST50977.2020.00065.