Optimizing Keywords in Financial Job Descriptions using Salary as the Target Variable

ECO421: Machine Learning in Finance
Professor Marlene Koffi

Yifan Zhu: zhuyif12, 1006345849 yvonneyifan.zhu@mail.utoronto.ca
Li-Wei Wang: wangliw4, 1004224724 jamiewang.wang@mail.utoronto.ca
Dorje Kongtsa: kongtsad, 1004516570 dorje.kongtsa@mail.utoronto.ca

Key words: machine learning, NLTK, tagger, K-mean, Naive Bayes Classifier, TF-IDF, labour market, financial industry

1. Introduction

In a competitive financial industry, hiring the right talent is a priority since finding the right skill sets for a vacancy significantly determines how much value can be created from the candidate. When a job vacancy is available, corporations often take immediate action to fill in the gap. Hence, writing a suggestive job posting becomes rather not important. Many companies ignore the attention for job postings. As a consequence, a generic description of the vacancy may attract a number of irrelevant applicants. Therefore, the matching efficiency is positively impacted by the recruiter's choice of terms.

A number of research has been done on terminologies that may be discouraging such as jargon, clichés, and unwelcoming words. However, salary information is also a quantitative feature in determining the attractiveness of an application. Exhibiting salary ranges in a posting can certainly give recruiters a competitive edge in attracting candidates that are concerned with a position's compensation and benefits. A recent study carried out by the job-hunting site Glassdoor.com demonstrates that money is the top motivator for 67% of job applicants (Maurer 2019). The vice president of a recruitment marketing agency, HireClix, stated: “Most people have choices in this job market. They don't want to waste their time with a role where pay and benefits don't meet what they're after”. Applicants will decide if the salary is not aligned.

The goal of our study is to develop a profile for the financial industry with the aim of determining the optimal keywords in a financial job description that can reflect salary expectations. In order to establish a complete profile, the study covers the general pattern in the financial job market and how the pattern identified and other factors reflect one's salary expectation. The objective is to investigate the predictability of job descriptions posted under the financial industry for salary expectations and detect indicative words that recruiters can optimize if it is not in their best interest to present the salary range. The project focuses on optimizing keywords in job descriptions by comparing the summarized job posting and the full detailed description; therefore, once a significant difference is observed, we will be able to conclude that recruiters lose potential candidates due to inaccuracy in summarization. Moreover, singular keywords can imply certain connotations; this paper aims to demonstrate that keywords selection based on their frequency in job descriptions can be helpful in making a judgment on the salary of the job. Collectively, we will be able to provide some comprehensive recommendations on the choices of words and phrases that will present a less salary-biased posting.

1.1 Literature Review

The similarities are in the methodologies used to process text data and extract/rank keywords. Specifically, cluster analysis is used in three different papers relating to text mining for job descriptions. The differences lie in the context of job descriptions and the purpose of extraction. For example, in the paper on "Text Mining for Industry 4.0", the job descriptions are mainly derived from the manufacturing industry whereas in our project we only analyse the financial industry (Pejic-Bach 2018). Another example of difference is "Job Description Mining for Work Integrated Learning" where authors use a time series analysis across a decade to understand the evolution of key words relating to the IT sector (Chopra 2018). The direction of research for the literature relies on the context of the job. In the Industry 4.0 paper, further research is directed to managerial roles in utilizing key words from manufacturing related jobs. In our project, the direction is to improve the selection of keywords in financial job descriptions.

2. Data and Methodology

2.1 Data

To investigate the predictability of job descriptions on current salary expectation of the job offered in the financial industry, this study uses Indeed.com as a source for analysing job postings, specifically, we are interested in jobs in Ontario, Canada. Two datasets are used in our work, both of which are generated through web-scraping via Python. The first dataset aims to analyse job postings shown on the result page, with "Finance" and "Ontario" as keywords, that are immediately observed by applicants searching on Indeed.com. In order to assess the reflectivity of the brief summary on the salary level, we anchor other observed variables as references. After removing replicate data and applying basic data cleaning techniques, the first dataset we obtained on October 28, 2021 contains 1039 observations with 8 variables: *Job title, Company, City Location, Multiple Locations, Remote, Post (*posting salary or not*), Salary (*average salary per hour*), Brief Summary (*summarized job description*)*

The second dataset contains the full description of each job posting with salary posted, which were retrieved on November 28, 2021 via Indeed API. We use the dollar sign to detect whether the posting contains salary information. After removing the unwanted observations (i.e., use $ to indicate the amount of assets held by the company in the job description) and subtract the salary level, there are 89 observations with two variables in our second dataset: *Full description* (qualitative variable with unlimited length), *Salary* (quantitative variable)

2.2 Methodology

Our methodology consists of two major parts: pattern mining methods and model-based methods. In the first part, we are able to detect the word frequencies, patterns and topics using WordCloud visualization, tagger (NLTK package), and K-means clustering. In the second part, we transform the response into binary outcomes 0-1, and use the decision tree algorithm to classify the situation when the job posting will include the salary level. Also, the Naive Bayes classifier will be presented to evaluate the salary expectation. By comparing the predictability of salary expectation using summarized description versus the full description, we are able to conclude whether the brief summary shown on the front page is informative to the viewers.

2.2.1 Pattern Mining Methods: Feature Extraction

The first step we took on each dataset is pre-processing the job description, which includes the process of tokenization, case conversion, lemmatization and exclusion. The exclusion process is based on the stopwords provided by the NLTK package with additional words selected by researchers. For example, we added 'finance' and 'financial' to the exclusion list, since they will not provide any further information to applicants. In the second step, the most frequent words appeared were extracted and visualized via WordCloud. We are able to see directly which job position has higher demand under the current marketplace. Also, we are able to see the quality of the candidate that is valued more. In the third step, we used the tagger to learn the word classes in job posting with and without salary being posted. This is to investigate if there is a significant difference.

In the fourth step, we used the unsupervised machine learning algorithm 'K-means' to cluster the words used in job descriptions into certain topics. We chose K-means over the Latent Dirichlet Allocation (LDA) since the LDA model's performance is limited on the short document. Meanwhile, K-means can guarantee the convergence, and it produces tighter clusters. We determined the optimal K cluster by considering the elbow rule and the interpretability of the model. Below is the mathematical representation of K-means:

$$min\,\{m_k\},\ \{r^{(n)}\} \sum_{n=1}^{N}\sum_{k=1}^{K} r_k^{(n)} \left|\left|m_k - x^{(n)}\right|\right|^2$$

where $r_k^{(n)} = I[x^{(n)}\ is\ in\ cluster\ k\ ]$, $m_1, m_2 \ldots, m_k$ the center of each cluster

2.2.2 Model-based Methods: Salary Prediction

Our primary goal is to evaluate the predictability of a brief summary on the salary expectation; thus, we implemented a Naive Bayes classifier. This requires the target variable to be binary. Therefore, we assigned 1 to those above the average salary level of the financial industry in Ontario ($87,657 in local currency), and 0 otherwise. In order to determine if the job summary is informative to predict salary expectation, we first conducted the Naive Bayes model on variables except the job summary as the reference level. Then, we conducted another Naive Bayes model on the job summary only. If the model accuracy increases significantly, we can conclude that the brief summary applicants see on indeed.com can convey more information to the salary expectation.

Next, we repeated the above step but used the full description in the second dataset to build the classifier. As a result, we are able to compare the informativeness of the brief summary and the full description. If the model accuracy is raised significantly, we could conclude that summarized description as a preview is still lacking information. Moreover, we used the Column Term Frequency Inverse Document Frequency (TF-IDF) measurement to catch the top terms indicating the high salary and the low salary level. This statistical method is based on the ratio of the phrase appearing in a document (TF) and the ratio of the total number of documents that phrase appears over the total number of documents that exist in the collection (IDF). Thus, recruiters or the platform could further adjust their posts with specific indicative words to improve the recruiting efficiency to attract the target population, while the applicants can have more accurate expectations when salaries are not being posted.

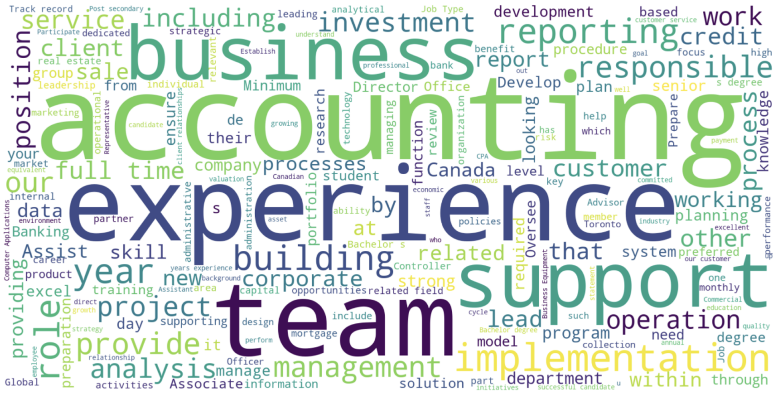
3. Results

We separate the study result into two sections: we first present the word frequency for each variable group via WordCloud and exhibit each subject through topic modeling; second, we demonstrate keywords that are suggestive towards salary level using Naive Bayes to estimate.

3.1 Word Frequencies and Pattern

We first seek to isolate the differences that posting a salary causes to the topics generated from a full job summary and the topics generated from the brief job summary. We begin at the WordCloud visualization and frequency table on job title and full job summary to find any patterns in the dataset. We have found the following highest frequency terms in detailed description as “accounting”, “business”, and “experience” among others which are presented in Figure 1. These high frequency terms reflect the jargon used in financial job descriptions; this helps us understand the scope of vocabulary that is used in these descriptions.

Figure 1: WordCloud visualization of words frequency in the brief summary



Second, the characteristics of the words used in the brief summaries that have salaries are similar to the characteristics in the brief summaries without salaries. The result presents the distribution of frequency are similar in that they follow NN having the largest frequency and verb gerund having the least frequency; thus, the composition of a job description does not change when salary is presented in Additionally, the characteristics between brief summary with salary and full description with salary are also indifferent which illustrates how recruiters adopt similar paragraph structure in terms of parts of speech in: brief summary or full summary, with salary or without salary.

After testing and comparing the mean squared error produced by different levels of K, following the elbow rule, the value kinks at K=4; hence, we define the optimal K with four clusters. Table 1 demonstrates the four topics for brief summary and detailed description where we see that the topic from the brief summary gave us clusters with similar words relating to accounting and business.

Table 1: K-mean clustering with K = 4 on brief summary versus full description

Brief Summary Topics	Banking, Operations, Financial Experience, Administration
Full Description Topics	Management, Client Facing, Operations, Accounting

The brief summary and detailed job description share the topic of “Operations” which highlights the importance of operations related tasks in the financial industry. However, the brief summary has topics of banking and financial experience that communicates with the applicant’s job experience to make sure that it matches with the desirable experience that is reflected in the brief summary. The full summary has topics of client-facing and management which can specify the expectations of the job rather than appealing to the experience of the applicant. Intuitively, since in the full job summary, there is more room to expand on the actual responsibilities of the job; whereas, there is limited space in the brief summary; the selection of words are prioritized to match the applicant's experience.

3.2 Salary Prediction and Indicative Terminologies

In general, more than half of the postings scrapped do not include salary range and unsurprisingly, the absence rate is higher for relatively well known companies which can be attributed to the fact that they offer brand and culture more than financial compensation. With the word frequency result, we can predict whether a randomly selected job posting would have a salary range included by taking its brief summary as the variable. We constructed a Decision Tree model with an accuracy of 0.7461; however, the feature importance is less concentrated where the significance of the top determinant feature ‘oversee_preparation_report’ is only 0.042. In practice, those words and phrases would not elicit an applicant’s salary expectation; rather, they are just indications of the likelihood in which salary range is presented or not.

Table 2: Model Accuracy, AUC for different input variables

Results/ Source	Other Variables	Brief Summary	Detailed Description
Model Accuracy	0.6349	0.6984	0.8333
AUC	0.6376	0.6985	0.7

In Table 2, we constructed another classifier using Naive Bayes to predict whether a job has a higher salary level than the average Ontario wage rate of $26.75/hr. The first Naive Bayes classifier is formulated using all 7 variables except the brief summary and has an accuracy score of 0.6349 which is improved to 0.6984 if the brief summary is included. Hence, we can conclude that the brief summary is to some extent meaningful although it does not produce a significant difference. We then derived a set of words that indicate the level of salary using TfidfTransformer by calculating the frequencies and probability of word appearance and assigning it to either Class 1 which indicates high salary or class 0 which indicates low salary. Table 3 displays the top five terms that are associated with each group.

Table 3: Top five indicative words for low and high level of salary
using brief summary and detailed description

	Low Salary Indicators	High Salary Indicators	Top 10 Frequent Words
Brief Summary	Accounting, Service, Prepare, Experience, Year	Implement, Oversee, Policy, Report, Analysis	Accounting, Business, Experience, Work, Implementation, Prepare, Oversee, Policy, Service, Building
Detailed Description	Work, Salary, Job, Management, Hour	Skill, Team, Data, Area, Ability	

If we, instead, employ the detailed description as our variables to predict salary, we see a significant improvement in the accuracy for Naive Bayes model with an accuracy rate of 0.833. The detailed descriptions yield a higher rate of correct prediction using the words within the posting meaning that brief summary is less concluded and informative; nonetheless, the set of data we have for full description is very limited in quantity; therefore, words derived are highly biased and should not be taken into account when answering our research question. Instead, it should be treated merely as a reference level to compare the model's accuracy.

4. Conclusion

To answer our research question: "What are the optimal keywords in a financial job description that can reflect salary expectation?", we analyzed a general pattern in the job posting and found that there is a difference in word classes and topics between job posting with salary range and job posting without salary range. We built several Naive Bayes classifiers to predict words that indicate the presence of salary as well as group words into indicative clusters based on salary level, and examine the predictability of using the brief summary and the full description.

There are several limitations to our data and methodology such as filtered data size for the detailed description only containing around 100 observations due to the limit of using indeed API to scrap the data, which means the sample size may not be big enough to be representative. In addition, the optimal number of clusters selected for the K-mean clustering method is a difficult balance between interpretation and coherence score. Nonetheless, our study provides meaningful data that can be beneficial for users on job posting sites. According to our study result, applicants can look for words such as 'oversee' and 'report' when selecting jobs since those words indicate a salary range that is higher than the average. Whereas for the recruiters, using high-salary or low-salary indicative words to redesign the job posting assists in communicating the job vacancy better, especially for smaller companies who are more vulnerable to the disadvantage of the absence of salary range.

Future research based on our study can be conducted with the direction on job marketing, company image and compensation importance. Particularly, financial compensation is not the sole factor in determining applicants' willingness to apply; thus, further study on other variables can collaboratively reveal considerations that an applicant desires. The transparency aspect of the job market is also a potential area to study as matching efficiency will be highly improved.

Bibliography:

Maurer, Roy. “Salary Is Most Important Part of Job Ad.” *SHRM*, SHRM, 16 Aug. 2019, https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/salary-most-important-part-job-ad.aspx.

Chopra, S., & Golab, L. (2018, January 1). *Job description mining to understand work-integrated learning*. Retrieved November 3, 2021, from https://www.semanticscholar.org/paper/Job-Description-Mining-to-Understand-Learning-Chopra-Golab/b7b1ae9caeba5d88c5ff67a573295b4d595167a9.

Heidarysafa, M., Kowsari, K., Bashiri, M., & Brown, D. E. (2021). Towards a knowledge discovery framework for data science job market in the United States. *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1*, 875–887. https://doi.org/10.1007/978-3-030-89906-6_56

Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International Journal of Information Management*, *50*, 416–431. https://doi.org/10.1016/j.ijinfomgt.2019.07.014