Enhancing Job Market Transparency with Natural Language Processing

Pascual Eley, Maria Manna, Frank Song April 2024

Abstract

Job postings often exhibit inconsistencies between advertised experience levels, compensation ranges, and the actual requirements outlined in job descriptions. These discrepancies pose challenges for both job seekers and employers. To address this issue, we propose a solution that leverages Natural Language Processing (NLP) techniques for reviewing job postings. Our models analyze job description text and predict key job attributes: salary and experience level. By doing so, they enable early detection of potentially misrepresented jobs. We utilize transfer learning from pre-trained transformer models with additional deep neural network layers to realize significant performance improvement over a simple word2vec average embedding approach.

Introduction

In today's dynamic job market, digital job boards serve as pivotal platforms for individuals navigating employment opportunities and career transitions, particularly in sectors where frequent job changes are essential for skill development and growth. Despite their prevalence, digital job postings present challenges for both job seekers and employers. Misalignments between stated experience levels, compensation ranges, and actual job descriptions necessitate rigorous vetting by job seekers, leading to significant time and resource investment for all parties involved. The current prevalence of remote work and the globalization of job markets have compounded these challenges, necessitating more robust solutions for ensuring transparency and fairness in the hiring process.

Job postings often outline desired qualifications, including industry-specific certifications and skills, but a consistently stated requirement is the duration of experience in the relevant sector. Although other studies have concentrated on skill extraction from job listing text in order to more appropriately match candidates to jobs, we have yet to encounter any that evaluate the adequacy of the stated work experience requirements. In light of this gap, this paper proposes a solution leveraging Natural Language Processing (NLP) techniques to automate the review of job postings, aiming to identify discrepancies between advertised and actual job characteristics. By extracting and analyzing textual data from job descriptions, our models can predict key job attributes such as salary and required experience level, allowing early detection of ambiguous or misrepresented jobs.

We develop a multi-task predictor model that evaluates the primary job description text against its advertised compensation and experience requirements. This empowers both job seekers and posters to identify inconsistencies or mismatches between the attributes advertised in job postings and the actual job requirements early in the search process. Ultimately, our solution enhances the effectiveness of job boards, streamlining the job application process. Employers save time by avoiding interviews with unqualified candidates who misunderstand experience requirements, while job seekers save time by focusing on positions that align with their compensation expectations.

This paper leverages three models, using $\text{BERT}_{\text{base}}$, BERT Multilingual, XLM-RoBERTa, and JobBERT to evaluate thousands of global job listings across various industries and languages. We find that these models successfully detect instances where job listings specify required experience or compensation levels that do not align accurately with the actual job requirements. These results suggest a promising trajectory towards implementing more automated reviews of listings on digital job boards.

Background

Beginning in 2019, there was a notable increase in research efforts aimed at tackling classification challenges linked to digital job postings. The primary focus has been on skill classification, aiming to extract both hard and soft skill competencies from job postings [1, 2, 3]. Predominantly, researchers have relied on pre-trained BERT models for this task, which receive relatively high accuracy [1, 2, 3]. While most of these studies focus on job postings in a single language, Zhang et al. (2022, 1) bridged the gap into multilingual models. They employed cross-lingual transfer learning, utilizing English job postings to classify skills in Danish jobs postings. They find that BERT, JobBERT, and RemBERT all result in similar macro-F1 scores on English job listings, but that RemBERT significantly outperforms other models for Danish listings, likely due to its pre-training on many languages, to include Danish [1].

Zhang et al. (2022, 2) repeats this process for English-language job postings. Here, they test domain-adapted models, such as JobBERT and JobSpanBERT, against non-adapted models, concluding that the tested domain-adapted models exhibit superior performance[2]. They also test single and multi-task learning with their models, and report that single-task learning yielded better results. Drawing inspiration from their findings, we incorporate the JobBERT model as one of our primary models of interest in this paper.

Additional related work has been done in the field of salary prediction. Despite the recent trend to include salary data on more job listings, such listings are still a minority, therefore research into the field of salary prediction is somewhat limited [4]. Nonetheless, Bana (2022) has made notable strides by employing pre-trained BERT embeddings to forecast salaries using job postings scraped on popular job board websites. Bana's approach involves inputting the listing text into a BERT layer, followed by utilization of a Convolutional Neural Network (CNN) to summarize the posting. This summarized job description is then fed into a max pooling layer to capture the most pertinent features from each sentence. Finally, after flattening and passing through a dense layer, the model predicts a single salary value [5]. This work shows the benefits of using pre-trained BERT embeddings, recognizing as we do that the max token length of 512 tokens is a limitation to this model [5].

Prior to this, Chen et al. (2020) performed salary prediction for jobs in England using a Graph Convolutional Network (GCN). While they acknowledge that the GCN offers a simplified approach to solving the salary prediction problem, their work serves as a valuable reference point. Chen et al. utilized salary buckets (referred to as salary categories in their work) and employed a broader definition of required experience, analyzing certifications, other stated competencies, and time-based measurements of past experience. Notably, the data they utilized was limited to a relatively small geographic area, which minimized variance due to other factors and yielded decent accuracy in salary prediction [6].

Our research diverges from existing approaches in four main ways: first, we refrain from assigning value to specific competencies or certifications. Instead, we concentrate on comparing the advertised experience requirements of positions within the model’s knowledge base and assign a market value based on the description text of the job listing. Second, we adopt a multi-task approach that encompasses both salary and experience level prediction from the text. We do not omit data from any sectors; we use a robust multi-industry dataset that includes full and part time jobs, remote and in person jobs, and jobs with various pay scales (hourly, weekly, and annually). Lastly, we utilize a multi-language model that is both trained and tested on a variety of languages.

This approach serves as a robust market-checking tool, aiding job seekers in identifying suitable positions based on their experience level and evaluating individual listings for potential compensation ranges. It also empowers job posters to reassess selected position characteristics in response to the model’s feedback. By evaluating the appropriateness of time spent in the respective job field alongside compensation, our model offers a comprehensive assessment of job postings, bridging a significant gap in existing research.

Methods

Data

Our research utilizes two distinct datasets to construct a comprehensive corpus of job postings: the "US Job Postings from 2023-05-05" dataset and the "Job Posting Dataset for Luxembourg (LU)." The former comprised 33,000 job postings scraped from various US platforms, while the latter provided about 215,000 job postings scraped from predominant Luxembourg job platforms. One of the primary challenges encountered during data preprocessing pertained to the variability in the format and presentation of job attributes, particularly required experience and forecasted compensation. To address this, we adopted a standardized approach to categorize job attributes:

Experience Level Categorization: Diversity in the articulation of required experience necessitated the creation of six distinct experience levels: student/intern, entry, junior, mid, senior, and executive (Table A, Appendix). Each level is associated with a predetermined amount of experience expressed in time units, such as months or years, or was associated with specific keywords which indicated the appropriate level, such as “CEO.” Wherever possible, missing experience values were inferred from job titles or textual job descriptions. Instances where the discrepancy between stated and inferred experience levels was irreconcilable were flagged and excluded from further analysis. If advertised experience requirements were only stated in terms of education (technical school, doctorate degree, etc.) we excluded the job from our final dataset as we could not determine if experience was required, and if so how much. This constituted approximately 6% of job listings.

Salary Standardization: Salary values exhibited considerable heterogeneity in terms of format, currency, and frequency (hourly, weekly, monthly, and annual). Anomalous salary entries, such as outliers or nonsensical values, were identified and either corrected or excluded from the dataset. If a currency was not explicitly stated, we assumed all compensation data from the US job postings dataset was in USD, and all data from the Luxembourg dataset was in EUR. To ensure comparability and facilitate analysis, all salary values were converted to annual USD equivalents at the conversion rate of 1.1. Subsequently, salary values were grouped into predefined buckets to simplify interpretation and analysis. These buckets were adjusted as we fine tuned model parameters (Table B, Appendix).

Listings exhibiting significant discrepancies between stated and inferred attributes were flagged for manual review and potential exclusion. After consideration, we chose not to identify and remove duplicates, as we found multiple jobs with the same title could be listed by the same company with different compensation or text, and wanted our dataset to accurately represent all available listings. The final combined dataset resulted in a corpus of 93,747 job postings spanning 10 languages, predominantly English, French, and German. The data was over/under sampled to account for class imbalance.

Features

Our primary focus was on job description text. As such, we leverage the raw text of the job description as the primary feature for analysis. This text captures key information to include job responsibilities, required skills, and experience level. While we acknowledge there are potential benefits to incorporating features such as industry and location, we choose to utilize a simplified model aiming to streamline the analysis process and prioritize the most relevant features for our analysis.

Our models utilize pre-trained Tokenizers to include $\text{BERT}_{\text{base}}$, XLM-RoBERTa, $\text{BERT}_{\text{base}}$ Multilingual-Cased, and Whitespace Tokenizer. For our baseline model, we leverage word2vec. We limit the maximum sequence lengths to 512 tokens due to the input token limitation of BERT (and derivative pretrained models). We find that a notable portion of job descriptions in the dataset were longer than 512 tokens. However, we deemed the maximum token size of 512, approximately 384 words on average, to be sufficiently large to capture pertinent information while keeping compute requirements feasible [5].

Model Architecture

Baseline: We utilized a DAN architecture for our baseline. To create our DAN model, we utilized word2vec to process textual job descriptions and generate embeddings. The baseline model used initial salary buckets of $30,000, as displayed in Table B, Appendix. The initial baseline testing was performed with two identical models, each separately trained for level and prediction tasks. We acknowledge this may give the baseline models an advantage over our multi-task classification architecture, but still observed significant improvements when moving to transfer learning with BERT.

Model 1: After reviewing the performance of our baseline DAN model, we implemented two significant changes for Model 1. First we built the model architecture over a pre-trained $\text{BERT}_{\text{base}}$ model, with a 100-wide by 1-deep fully connected layer (with dropout) on top of CLS token output, followed by a split to separate classification heads, each with a 100-wide by 1-deep fully connected layer and softmax classification layer at the top. Second, we tested different salary buckets with the goal of utilizing buckets that were more appropriate for the task at hand (Figures B & C, Appendix). In place of the eight $30,000 intervals we used in the baseline model, we instead use seven uneven intervals with the intent of aligning the buckets to similar job-types so that it was a more intuitive model (Table B, Appendix). It is at this point in our experiment that we begin testing different pre-trained models, to include $\text{BERT}_{\text{base,}}$ BERT Multilingual, and XLM-RoBERTa.

Model 2: For our second model, based on the distribution of false positives for each salary bucket, we consolidated from 7 salary buckets to 5 (Table B; Figures D & E, Appendix). Two additional changes were made for the second model: first, all hidden layers had width increased to 200 and depth increased to 2, and second, the learning rate was decreased from 0.00005 to 0.00001 for our XLM-RoBERTa based model.

Model 3: Next we employ a model that is more closely aligned with the contextual text data we are investigating in this paper (Figures F & G, Appendix). Model 3 leverages the JobBERT model where we expect to see an improvement vs $\text{BERT}_{\text{base}}$, as it is continuously pre-trained from a $\text{BERT}_{\text{base}}$ checkpoint on ~3.2M sentences from job postings. The learning rate was adjusted back to 0.00005.

Model Training

We split our data into a single train/validation/test set that was used across all models. Our baseline DAN model was trained on a T4 GPU, and our transformer-based models were trained for 3-5 epochs each on an A100 at batch size 32 (~30 min/epoch). Batch size was chosen based on previous results by Gnehm et al. 2022, and to balance training time with available GPU capacity.

Evaluation Metrics

Metrics of interest include precision, accuracy, recall, and macro-F1 scores. This combination of evaluation metrics ensures a comprehensive assessment of our model, tailored to the multi-task nature of the problem statement. Utilization of these metrics also allows a clear comparison between models internal to this study and the outside studies previously discussed.

Although we ultimately decided to utilize salary buckets as classes, performance seems to indicate that the model did in fact learn to place the categories along some ordinal spectrum. Many of the incorrect salary predictions are only one category removed from the true category (for example, 23.7% of examples with a “true” label of 100-150k are predicted to be either 50-100k or 150-200k). Although we shrank the salary buckets to improve performance on our chosen evaluation metrics, we might choose to operationalize the salary prediction loss and evaluation differently to take advantage of the ordinal nature of salary labels in future studies. Other authors have seen some success using RMSE, which may help our dual output model overcome the salary bucket threshold issue [5].

Results & Discussion

The outcomes of our experiments are presented in Table 1, detailing the performance metrics of each model in terms of the weighted average accuracy, precision, recall, and F1 scores. Although disparities among the pre-trained models are generally minimal, a clear frontrunner emerges. Compared to our baseline, we observe significant enhancements with Model 3. Model 3 (JobBERT) demonstrates substantial improvements in accuracy and F1-score compared to Model 1 (BERT$_{\text{base}}$ and BERT Multilingual). This performance boost can be attributed to JobBERT's domain-specific pre-training, which, despite being trained on a single language, appears to provide a significant advantage in understanding and processing job-related text data [2].

In contrast, Model 1 (BERT base and BERT Multilingual) exhibit comparable performance to the baseline. While we expected BERT Multilingual to perform better due to its pre-training on multiple languages, including the primary languages in our dataset, the marginal improvements observed over the baseline suggest that the multi-lingual aspect may not significantly impact model performance in our specific task.
Model 2's utilization of XLM-RoBERTa, also aimed at leveraging its multilingual pre-training, again results in comparable performance to BERT$_{\text{base}}$ [7]. This reiterates that while multilingual pre-training can provide benefits, it may not always translate to substantial performance gains when text is in common Western European languages present in the BERT$_{\text{base}}$ pre-training corpus.

Our model adjustments to enhance performance reflect a strategic approach to leveraging pre-trained models tailored to our dataset characteristics. The adoption of JobBERT in Model 3 underscores the importance of domain-specific pre-training in achieving superior performance in tasks such as job classification. Conversely, the utilization of multilingual pre-trained models in Models 1 and 2, while promising, may not offer significant advantages over baseline performance in our specific context.

Table 1: Model Performance

	Level / Salary (**BEST** / *WORST*)			
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Baseline	*0.31* / 0.48	*0.53* / 0.53	*0.31* / 0.48	*0.23* / 0.49
Model 1: BERT Base	0.66 / 0.54	0.74 / 0.58	0.66 / 0.54	0.65 / 0.55
Model 1: BERT Multilingual	0.60 / *0.43*	0.70 / *0.49*	0.60 / *0.43*	0.60 / *0.44*
Model 2: BERT Base	0.66 / 0.70	0.73 / 0.71	0.66 / 0.70	0.65 / 0.70
Model 2: XLM-RoBERTa	0.67 / 0.69	0.74 / 0.71	0.67 / 0.69	0.67 / 0.69
Model 3: JobBERT	**0.70 / 0.73**	**0.76 / 0.75**	**0.70 / 0.73**	**0.70** / **0.73**

**For metrics calculation details, please reference Appendix Note A*

Upon closer examination of our models' errors, we identify four main themes. First, subtle variations in job description phrasing occasionally resulted in misclassification of edge cases into adjacent categories, impacting both experience level and salary. Specifically concerning salary values, which we converted into ordinal salary buckets but were originally either single numbers on a continuous pay scale or a given compensation range, we observe that the discrepancy between the true value and the assigned bucket is typically minimal, often just a few thousand dollars. In these cases, while a strict interpretation might suggest our model's misprediction of the salary bucket, we contend that the transformation of continuous variables to ordinal ones introduces inherent ambiguity.

Second, we find that our assumptions regarding experience levels may not universally apply across industries. For instance, while a financial services sector standard may stipulate a Senior Accountant requires at least five years of prior experience, our model's classification of two to five years as a "mid" level job may conflict with original labels. This reveals a potential flaw in our model's understanding, stemming from assumptions about experience level definitions.

Moreover, we note that job descriptions emphasizing company importance or proximity to key figures tend to be classified as higher levels, indicating a bias in our model towards certain contextual cues. We compared predictions from our best model for a sample set of job postings from LinkedIn, where we selectively included portions of the posting with: 1.) only the job duties and skills, 2.) additional company context information, and 3.) additional company information that was synthetically enhanced by GPT-4 (Table D, Appendix). Our findings support our hypothesis about how our model's results are impacted by this additional information commonly included in postings.

Lastly, we encounter instances where our model's predictions diverge from true labels due to misleading or inaccurate job description metadata. Despite this discrepancy, upon closer examination, our model often correctly identifies mismatches between advertised and actual data. For example, a job description originally in French claimed a compensation exceeding $100,000, yet a translated text revealed inaccuracies in salary and job level data, leading to a misrepresentation. Despite the inflated figure, our model correctly inferred a lower salary bracket, highlighting its effectiveness in flagging discrepancies for further review. This analysis underscores our model's efficacy in addressing our problem statement, as it appropriately identified discrepancies and misleading information within job descriptions.

Conclusions

When we started this work, we endeavored to accurately predict both the salary range and job level of a position given only the text from its description. While a simple word2vec averaging approach showed an ability to provide better than random predictions, utilizing transfer learning with pre-trained transformer models led to significantly better results. We ultimately were able to reach good levels of accuracy, and even observed that incorrect predictions tended to be close to the true label on the inherent ordinal scale of the labels we're working with. Some texts were harder for the model to decipher, which our analysis suggests was due to context language about the company, clients, or mission that is similar language included in responsibilities for higher job levels. To address both of these findings, we'd likely use future research opportunities to operationalize our salary predictions in a different way, and to experiment with a two stage approach to detect/extract context information from job postings before trying to predict our desired output.

References

1. M. Zhang, K. Nørgaard Jensen, B. Plank, *KOMPETENCER: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning;* Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 436–447, Marseille, France, June 20-25, 2022
2. M. Zhang, K. Nørgaard Jensen, S. Dam Sonniks, B. Plank, *SKILLSPAN: Hard and Soft Skill Extraction from English Job Postings;* Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4962 - 4984, July 10-15, 2022
3. A-S. Gnehm, E. Bühlmann, H. Buchs, S. Clematide, *Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads*; Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 14 - 24, November 7, 2022
4. C. Stahle, *Pay Transparency in Job Postings Has More than Doubled Since 2020*; Hiring Lab, Economic Research by Indeed, March 14, 2023
5. S. Bana, *work2vec: Using Language Models to Understand Wage Premia*; December 20, 2022
6. L. Chen, Y. Sun, P. Thakuriah, *Modelling and Predicting Individual Salaries in United Kingdom with Graph Convolutional Network*; In: Madureira, A., Abraham, A., Gandhi, N., Varela, M. (eds) Hybrid Intelligent Systems. HIS 2018. Advances in Intelligent Systems and Computing, vol 923. Springer, Cham
7. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, *Unsupervised Cross-lingual Representation Learning at Scale*; 2020
8. Y. Sun, F. Zhuang, H. Zhu, Q. Zhang, Q. He, H. Xiong; Market-Oriented Job Skill Valuation with Cooperative Composition Neural Network; Nature Communications 12, Article No. 1992, 2021
9. Z. Wang, S. Sugaya, D. P.T. Nguyen, *Salary Prediction using Bidirectional-GRU-CNN Model*; 2019, The Association for Natural Language Processing
10. J. Devlin, M-W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*; May 24, 2020
11. S. Baccianella, A. Esuli, F. Sebastiani, *Evaluation Measures for Ordinal Regression*; 2009 Ninth International Conference on Intelligent Systems Design and Applications, January 2009
12. E. Amigo, J. Gonzalo, S. Mizzaro, J. Carrillo-de-Albornoz, *An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results*; Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3938–3949, July 5-10, 2020
13. K. Miller, *Words Matter: The Text of Online Job Postings Can Predict Salaries*; Stanford University Human-Centered Artificial Intelligence (HAI), January 10, 2022

Appendix

Table A: Experience Level Definitions

Experience Label	**Definition**
Student_Intern	Explicitly Stated
Entry	No prior experience
Junior	0-2 years
Mid	2-7 years
Senior	7+ years
Executive	Explicitly Stated

Table B: Salary Buckets

A (Baseline)	**B (Model 1)**	**C (Models 2, 3)**
<$30,000	<$45,000	<$50,000
$30,000 - $60,000	$45,000 - $65,000	$50,000 - $100,000
$60,000 - $90,000	$65,000 - $85,000	$100,000 - $150,000
$90,000 - $120,000	$85,000 - $110,000	$150,000 - $200,000
$120,000 - $150,000	$110,000 - $150,000	>$200,000
$150,000 - $180,000	$150,000 - $200,000	
$180,000 - $210,000	>$200,000	
>$210,000		

Table C: Model Parameters

	Baseline	**Model 1**	**Model 2**	**Model 3**
Hidden Width	[100,50]	100	200	200
Hidden Depth	2	1	2	2
Learning Rate	0.001	0.00005	0.00005 / 0.00001	0.00005
Pretrained	N/A	BERT-base / BERT-multilingual	BERT-base / XLM-RoBERTa	Jobbert
Epochs	20	3	5	5

Table D: Analysis of Actual Intern-Level LinkedIn Job Posting with JobBERT, Model 3

	Job Description Only	**Job Description & Company Description**	**Job Description & Synthetic (GPT-4) Enhanced Company Description**
Predicted Level	student_intern	student_intern	mid
Predicted Salary Bucket	<$50,000	$50,000 - $100,000	$100,000 - $150,000

Figure A: Distribution of Job Description Text Lengths

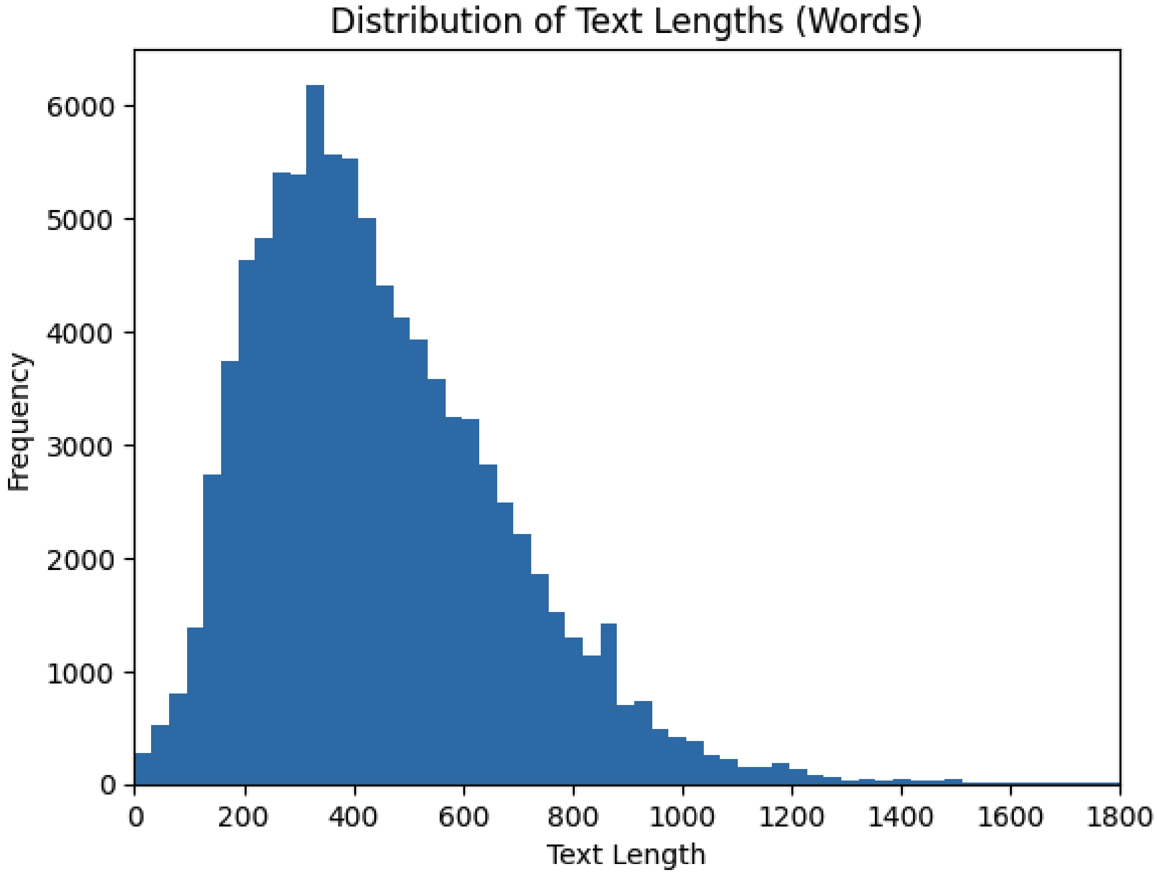


Figure B: Model 1, BERT$_{base}$ Confusion Matrix – Level

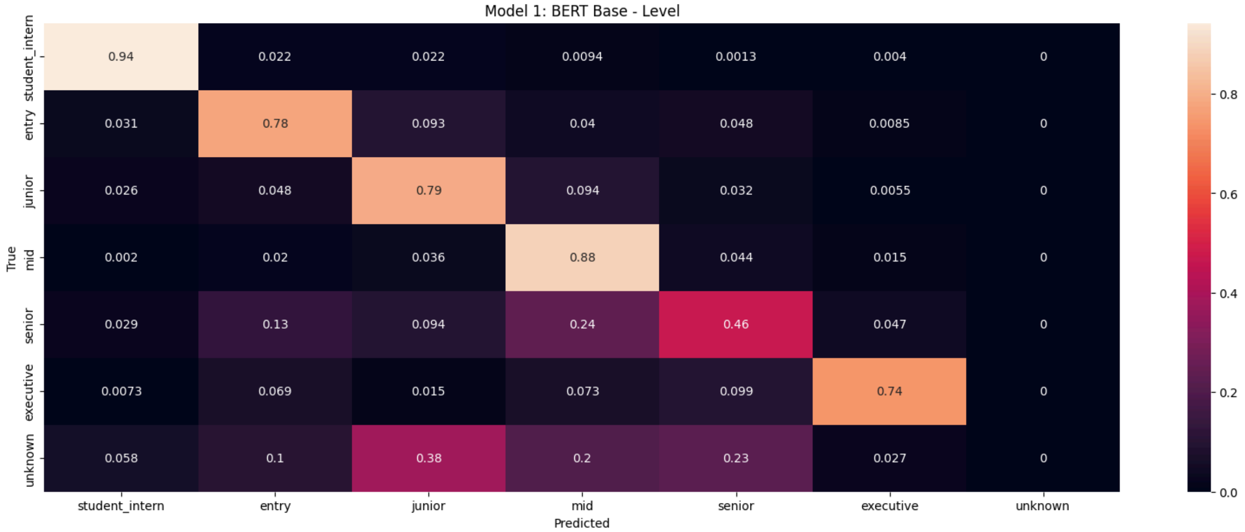


Figure C: Model 1, BERT$_{base}$ Confusion Matrix – Salary

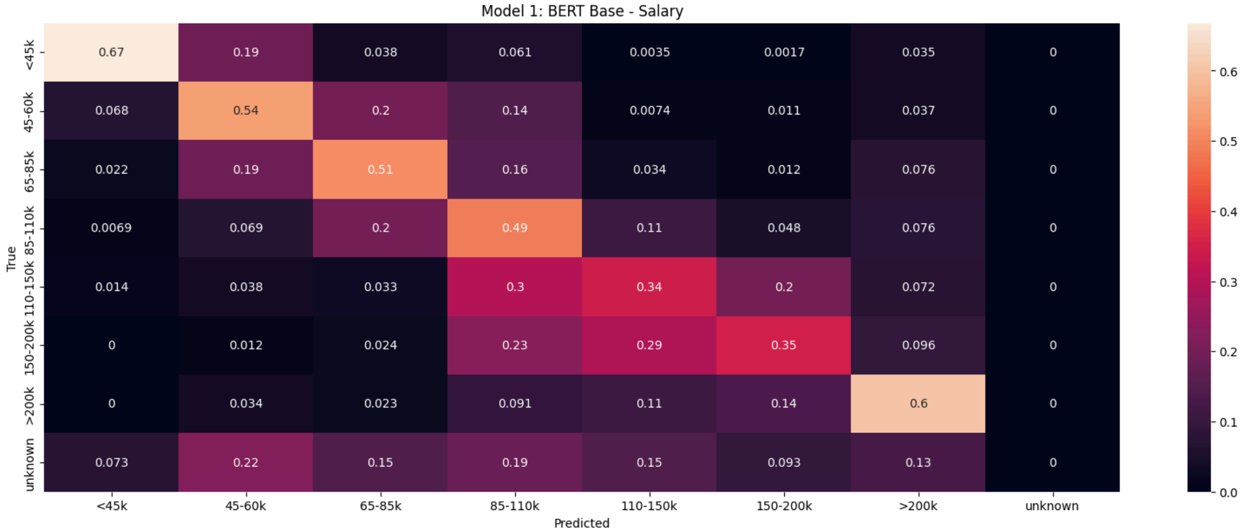


Figure D: Model 2, BERT$_{base}$ Confusion Matrix – Level

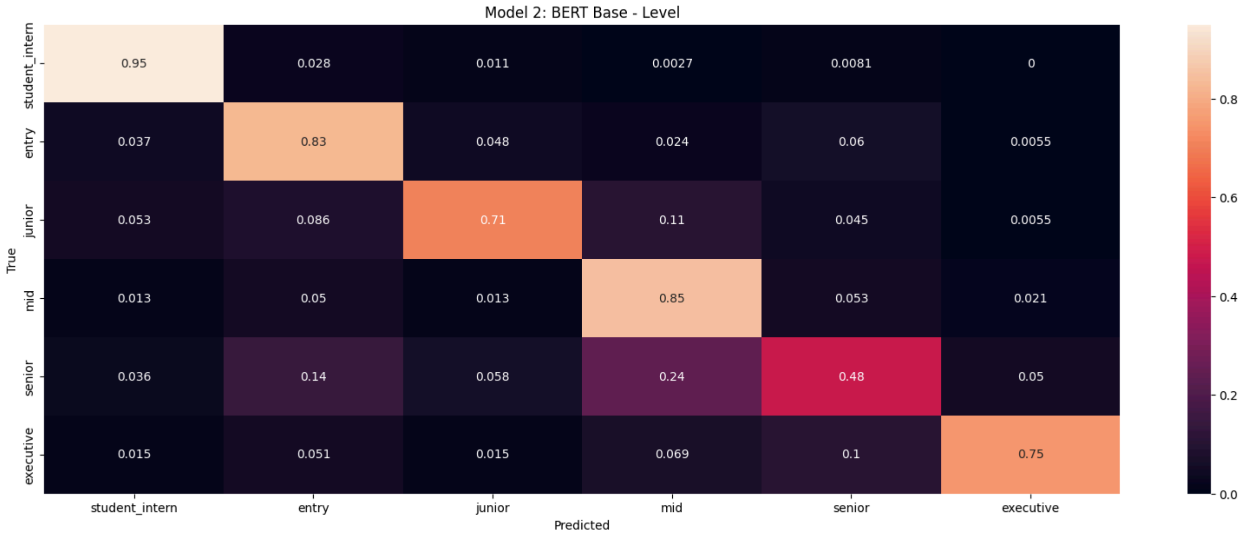


Figure E: Model 2, BERT$_{base}$ Confusion Matrix – Salary

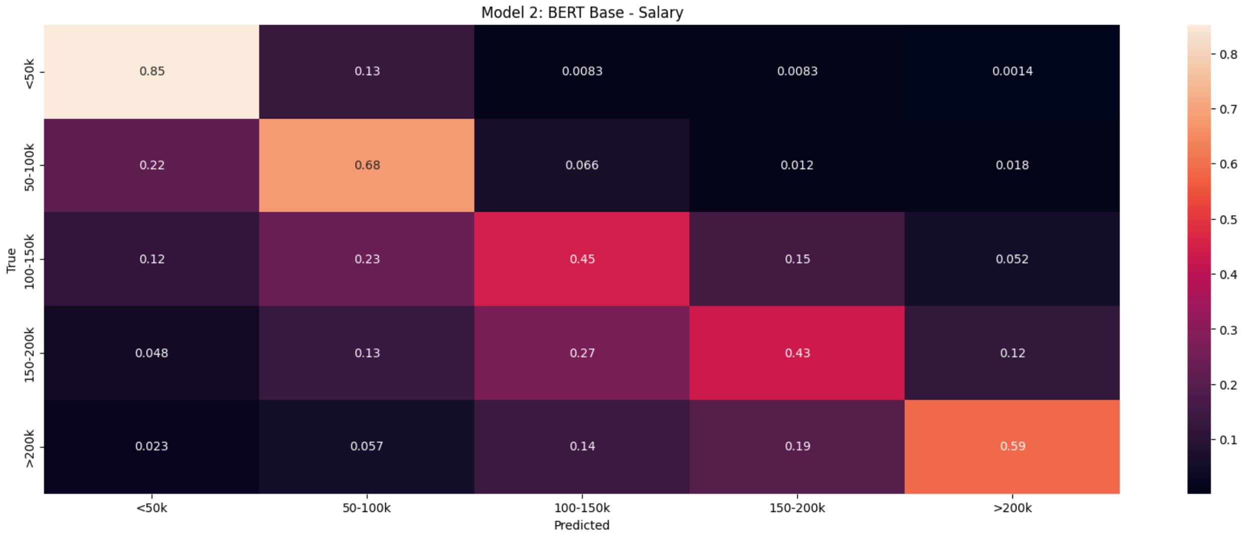


Figure F: Model 3, JobBERT Confusion Matrix – Level

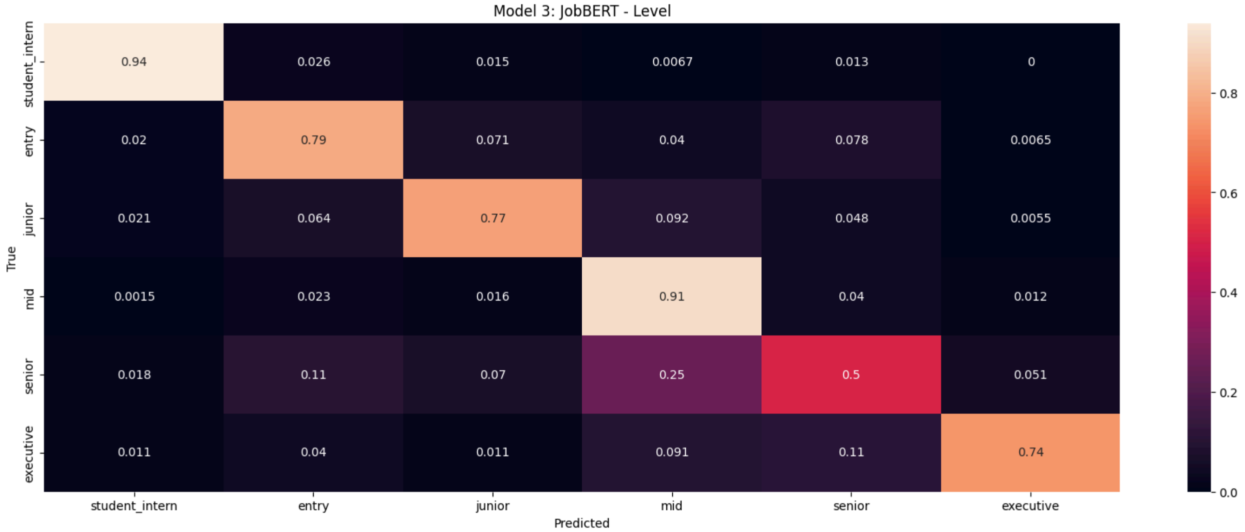
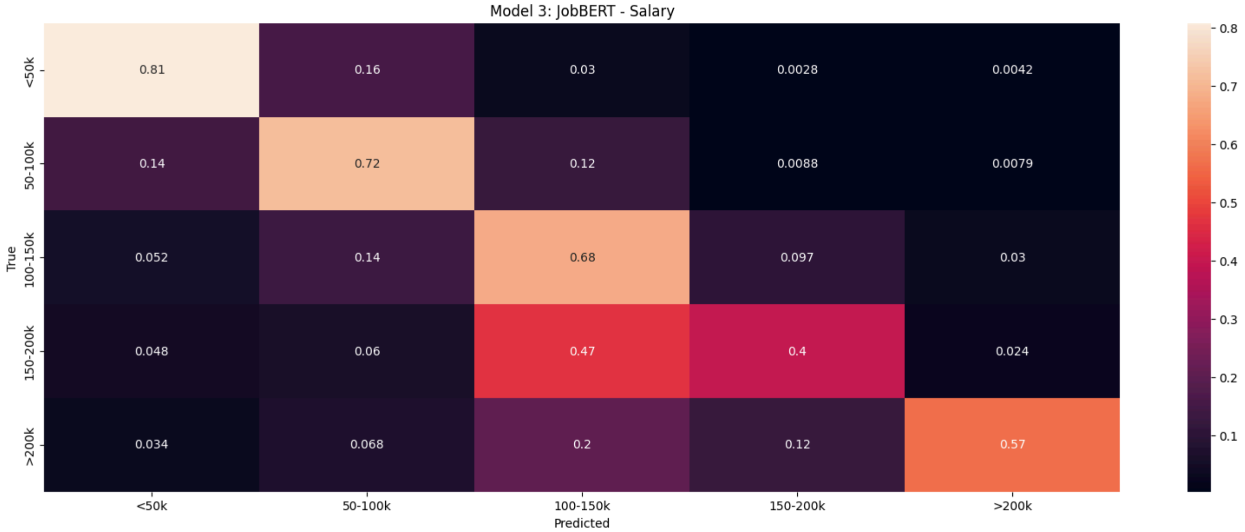


Figure G: Model 3, JobBERT Confusion Matrix – Salary



Note A: Filtering Test Set for Metrics Calculations

When calculating metrics for each classification type (salary and level) we filtered our test set to just examples with valid labels for the given classification. Because of our masked loss function, our model never “learned” to predict that a value was missing. Therefore, it will always predict one of the valid class labels. If we had not filtered the test set for each evaluation, the model would lose a significant amount of accuracy from “incorrect” predictions for examples with a true label of “missing”.