work2vec: Using Language Models to Understand Wage Premia

Sarah H. Bana*

December 20, 2022

Abstract

Does the text content of a job posting predict the salary offered for the role? There is ample evidence that even within an occupation, a job's skills and tasks affect the job's salary. Capturing this fine-grained information from postings can provide real-time insights on prices of various job characteristics. Using a new dataset from Greenwich.HR with salary information linked to posting data from Burning Glass Technologies, I apply natural language processing (NLP) techniques to build a model that predicts salaries from job posting text. This follows the rich tradition in the economics literature of estimating wage premia for various job characteristics by applying hedonic regression. My model explains 83 percent of the variation in salaries, 19 percent (13 percentage points) over a model with occupation by location fixed effects. Using an attribution method called integrated gradients, I decompose these elements into locations, job titles, experience levels, education levels, skills, activities, and firm names. This decomposition demonstrates the relative contribution of each of these factors to earnings.

*Chapman University and Stanford Digital Economy Lab. Email: sarah.bana@gmail.com. This is a companion paper to work by myself, Erik Brynjolfsson, Daniel Rock, and Sebastian Steffen entitled "job2vec: Learning a Representation of Jobs." Throughout the course of this project, I was told that "job2vec" was trademarked. Accordingly, we have adjusted to using "work2vec." I am very grateful for thoughtful comments from seminar participants at the Stanford Institute for Human-Centered Artificial Intelligence, Boston University, Emory University, Simon Fraser University, the University of Minnesota, the Brynjolfsson Lab, MIT Initiative on the Digital Economy, University of California Santa Barbara Applied Lunch, Chapman University, and the University of Nevada at Reno, along with countless friends and family members. Special thanks for the funding from the Stanford Institute for Human-Centered Artificial Intelligence's Google Cloud Credit Grant that enables this research. All errors are my own.

1 Introduction

It is estimated that 1.7 megabytes of data are created every second for every person on Earth (Domo, 2018). This is largely consistent with more and more of our activities being conducted online. One activity increasingly performed online is job search – firms post jobs and interact with candidates virtually, and workers use online job boards to identify opportunities (Kuhn and Mansour, 2014). The data produced from these activities has the potential to generate unprecedented insights into firms' production functions and workers' activities.

Such data is often unstructured. For example, job postings are free-form text of varying length, without well-defined fields. Thus far, researchers have condensed such text into structured data by identifying relevant key words or adding high dimensional fixed effects by categorizing jobs into discrete buckets.

The text is rich and new tools in computer science have demonstrated breakthrough performance in "understanding" text (Devlin et al., 2018). At the same time, a paradigm shift in artificial intelligence (AI) systems has led to the growth of foundational models – models that are trained on broad data at scale and can be adapted to a wide range of downstream tasks (Bommasani et al., 2021). This approach significantly reduces the computational cost of using text data. One foundational model, Bidirectional Encoder Representations from Transformers (BERT), provides context-dependent embeddings (dense vectors) that can readily be used as the first layer of a model.

With these tools, I train a natural language processing (NLP) model on the text of job postings, and demonstrate that the text of the posting matters. This model takes the text of the posting as an input and translates the text to vectors using BERT's pre-trained word embeddings. These word embeddings, for example, will produce different vectors for the word "models" when characterizing a job advertisement that states "deploy machine learning models" compared to "models exceptional customer service." This initial layer produces a matrix of 512 by 768 dimensions for each job posting. Additional model layers condense the dimensionality.

Using a new dataset with salary information from Greenwich.HR linked to posting data from Burning Glass Technologies, I can reframe salary prediction as a supervised learning

problem. My model, incorporating the text, explains 83 percent of the variation in salaries, a 19 percent (13 percentage point) increase over a baseline with occupation fixed effects by Metropolitan Statistical Area (MSA) fixed effects. On another relevant metric, the natural language processing model represents a 39.3 percent decrease in the out-of-sample Root Mean Square Error (RMSE).

The language model outperforms other supervised learning models taking into account the skill clusters tagged in the postings, suggesting that the context matters for salaries, and postings provide information about the job and its wage, above and beyond the skills requested.

Estimating wage premia for various job characteristics by applying hedonic regression has been common in the economics literature (Mincer, 1974; Heckman et al., 2006; Weinberger, 2014; Deming, 2017). Hedonic regression techniques uncover the predictive value of characteristics for equilibrium outcomes in the market. Because both sides of the market are heterogenous, the equilibrium prices provide information to both firms and workers.

This work builds on Autor and Handel (2013), Deming and Kahn (2018) and Marinescu and Wolthoff (2020), three pioneering papers that highlighted the wage heterogeneity within occupation and demonstrated that additional characteristics like tasks, skills demanded, and job titles can explain this variation.

Autor and Handel (2013) conduct a survey to collect new data on the job activities of a representative sample of U.S. workers across task domains, and demonstrate that within-occupation measures have significant and economically meaningful predictive power for earnings. This process relies on nationally representative survey data for a sample of 1,333 workers. The drawback of this approach is its lack of scale: to identify rare characteristics, the sample must be substantial. To identify differences over time, the survey must be conducted repeatedly.

Papers that followed used data from online job boards. The advantage of this approach is that these analyses can be done in closer to real-time, and avoid costly surveys. Deming and Kahn (2018) show that skill requirements affect average wages of professionals across MSAs, explaining up to 94% of the variation in average wages in MSA-occupation cells. The analysis focuses on average wages, when there is substantial variation *within* occupation in wages. Fur-

thermore, the sample is understandably limited to professional job advertisements, as during that time period (2010-2015), online job postings leaned heavily towards professional occupations. Marinescu and Wolthoff (2020) find a coefficient of determination (R^2) of almost 90%, looking at the explanatory power of job titles using posted wages on Career Builder. This number is remarkably high, but is limited to the sample of under 20% of postings that posted wages. Given that postings with and without wages systematically differ, this may be difficult to extrapolate to the general population. My work extends this research by introducing a new dataset with salaries derived from the metadata of job postings. I also demonstrate that job titles have little *out-of-sample* predictive power because the number of unique job titles is very high.

Differing from previous interpretable approaches like high-dimensional fixed effect regressions, the natural language processing methods used in this paper often lack interpretability. In the context of this research, this implies that though the model can explain what differentiates a high and low salary job posting, it is difficult to translate this information into actionable insights. However, related advances in explainable AI allow us to assign an importance score to each input feature. Integrated gradients is one popular attribution method, for its properties sensitivity and implementation invariance (Sundararajan et al., 2017).

The output of this method is an importance score for every token in a posting. I describe the most positive and negative attributions, which have interpretable magnitudes in terms of their contribution to log earnings. The most positive words are associated with job titles, such as "dentist," "psychiatrist," "superintendent," "director," and "engineer." On the other hand, the most negative words are associated with experience and role characteristics, such as "intern/internship," "cleaning," "entry," "hour," and "aide."

I then split words into categories to informally evaluate the model's output. Locations in coastal cities and California are associated positively with salary, while states like Florida, Kansas, Kentucky, Missouri, and cities like Tucson and Katy are associated negatively with salary. This approach is repeated with acitvities, technology skills, and job titles.

Using this model, I decompose the types of words used in job postings into categories that

have either a theoretical or empirical basis for affecting earnings. These categories are locations (Card et al., 2021), job titles (Marinescu and Wolthoff, 2020), experience levels, education levels (Mincer, 1974), skills Deming and Kahn (2018), activities (Acemoglu and Autor, 2011), and firm names (Bonhomme et al., 2020). Because there is no singular dataset that exists with all of these factors, little is known about the relative contribution of each of these factors to earnings.

[PLAN]

2 Data

The data comes from two distinct data vendors, Greenwich.HR and Burning Glass Technologies. In this section, I describe the elements of the data used for each portion of the analysis.

2.1 Greenwich.HR Data

Greenwich.HR (GHR) is a labor market intelligence firm that provides real-time labor market data to application developers, analysts and consultants. They consolidate job postings from millions of different sources. A major advantage of the GHR data is that they have collected pay data for over 70 percent of job postings collected in recent months. Though the exact method by which GHR collects this data is proprietary, I outline the approach in general terms to lend credence to the estimates.

While many postings do not contain information on wages, it is common practice for job posting platforms to solicit salary data from the recruiter posting the job. For example, in Figure A1 Panel A and B, it can be seen on one popular platform, Indeed, that recruiters are encouraged to fill in either the exact rate, the range, a starting salary, or a maximum salary. This screenshot is for illustrative purposes only, as the platforms and methods for integrating data used by GHR are proprietary. Panel A suggests that this incentivizes applicants. In Panel C, a similar screen is included for LinkedIn.

This information can be found on the applicant side when searching for postings. Visual-

ized in Figure A1 Panel D, a postings' salary band can be inferred by whether it appears in the search results when changing the pay threshold. These images are intentionally taken from different platforms to demonstrate the ubiquity of this practice.

Key for the analysis, the postings' salary band is drawn from the metadata of the posting, as opposed to the characteristics of the postings itself. That is, GHR does not create a mechanical correlation between the posting language and the salary reported.

This pay data provides a major asset for analysis. However, like many new datasets, there are limitations. First, GHR did not collect the raw job text until 2020. Second, GHR sought to be a comprehensive source of the U.S. economy only beginning in March 2019. Prior to this time period, the focus was on public firms and certain sectors. The first limitation can be overcome by connecting postings between Burning Glass Technologies, which does collect the full text of the posting, and GHR. The second limitation precludes time series analyses on the changing wage premia over time going back. Because the COVID19 pandemic occurred in 2020, likely changing the premia associated with certain skills, this work focuses on cross-sectional variation in wages during the year 2019.

GHR contains 62,026,448 job postings for the period April 2019 to September 2020 (18 months). Of these postings, 37,113,670 contain posted salaries (59.8 percent). The posting distribution is displayed in Figure 1. As evidenced by the jagged lines in the density distribution, posted salaries do bunch at round numbers.

2.1.1 Comparison to CPS

To the best of my knowledge, there is no source of nationally representative posted salaries to compare GHR data to determine potential selection issues. The best alternative is comparing the salary distribution to the distribution of weekly earnings in the Current Population Survey. The Current Population Survey (CPS) collects earnings from one-fourth of the monthly sample, limited to wage and salary workers. The closest comparison is usual weekly earnings, representing data before taxes and other deductions, and including any overtime pay, commission or tips usually received.

I use the fourth quarter in 2019's CPS release for this comparison, graphically depicted in Figure 2. The 25th percentile of CPS weekly earnings is $623, which at 52 weeks a year is $32396. This is quite close to the 25th percentile of GHR salaries, at $32175.19. The median CPS weekly value is $936, which is an annual value of $48,672. This is much lower than the GHR median of $41,750. This pattern continues, with the 75th percentile of CPS earnings is $77376 annually, while the GHR percentile is $66501.

There can be several reasons to expect the posting distribution and the actual salary distribution to differ. The two broad categories of reasons are (1) differences in job composition and (2) differences in reporting of pay.

The posting distribution represents new jobs, and therefore, industries and occupations that have higher turnover are likely to be overrepresented. For example, according to the BLS Job Openings and Labor Turnover Survey (JOLTS), the government sector has relatively low turnover, while the private sector has higher turnover. Within the private sector, there are also notable differences: leisure and hospitality is a high turnover industry, while durable good manufacturing is low turnover. Moreover, there are notable differences within occupations. In one extreme example, seasonal work has tremendous turnover, with large fractions of Lifeguards, ski patrol, and other recreational protective service workers being rehired at the beginning of every season. Given that higher turnover jobs are more likely to be lower wage, this is consistent with the overall directional difference between the posting distribution and the CPS distribution.

Differences in job composition between the posting distribution and the actual salary distribution can also be a function of how workers are hired. First, not all jobs are posted online. Previous research on online job postings has emphasized that as online job postings have become more common, firms and jobs added more recently are lower skilled (Blair and Deming, 2020). Moreover, not all jobs are posted and some postings may still represent more than one vacancy, despite the best attempts to deduplicate. To the best of my knowledge, there is no credible estimate of the fraction of jobs that are not posted, although ongoing work by researchers at the Bureau of Labor Statistics seeks to answer this question.

Though the job composition is likely different, the CPS and GHR are also measuring different underlying concepts. The CPS usual weekly wage includes expected overtime, commission and tips. These are not included in the GHR data.

The distributions are clearly different; however, it is difficult to assess whether this is a cause for concern. Future analyses will test robustness to various assumptions about the distribution.

2.1.2 Comparison between GHR Postings with and without Salaries

Another approach to assessing the representativeness of GHR salaries is to measure how much other observable characteristics can explain whether the salary exists. Using a 20 percent random subsample of postings from April 2019 to December 2019, I regress a binary for whether the salary is missing on six digit Standard Occupation Classification (SOC) code fixed effects. If occupations that are higher wage are less likely to be well represented with salary data, then occupation fixed effects should explain considerable variation in whether the salary is missing.

Instead, I find that the pseudo R^2 on a probit regression with occupation fixed effects is only 0.0134. That is, which postings have salaries in the data cannot be explained by the occupations of those postings. This is, by no means, conclusive evidence that salaries from metadata are random. However, it does suggest that the process by which salaries appear in metadata differs from what might be expected for posted salaries.

2.2 Burning Glass Technologies Data

Burning Glass Technologies (BGT) is an analytics software company that strives to provide real-time labor market information to higher education institutions, firms and municipalities. The product used in this analysis is the job postings data, collected from over 40,000 online job boards and company websites. These postings are deduplicated in a proprietary manner and the job title and employer name are cleaned.

For the analysis described, the key attribute of the data employed is the raw job text. This raw text has been seldom used in prior research, and contains virtually all the information that

the applicant will see. The job text frequently contains information about the firm, the role, and the application procedure, though this is not systematic.

For illustrative purposes, the raw job posting text of two sample postings from October 2019 are displayed in Figure 3. Both postings use different terms to convey similar information. For example, in the first posting, responsibilities are outlined in the "Key Responsibilities" section, while these same thoughts are outlined in the second posting under the heading, "What would you do? The Specifics." Postings also differ in length, and some postings have some information about benefits and how to apply.

I link a GHR posting with a BGT posting using the firm name, job title, and date of posting. The two datasets are cleaned differently, so connecting them involves a fuzzy match. Typos and extraneous information are more likely to be at the end of the firm name or cleaned title, which means a string distance measure that weighs the beginning of the string is preferred. For this purpose, I use a Jaro-Winkler distance metric.

3 Model

3.1 Model Structure

To describe the process by which the NLP model takes text as an input to predict salaries, it is helpful to think about the limits of traditional data for this analysis. Suppose our objective was to compare a group of job postings. We might transform this into traditional data by counting all the distinct words in each posting. The resulting matrix would be full of zeroes, as many postings would not contain certain words, creating challenges for traditional regression analysis. Moreover, the number of prepositions or conjunctions in each posting might not necessarily be meaningful. Even if the data was not sparse, simply counting words might be suboptimal: we improve the situation by counting pairs of words (called bigrams), instead of counting individual words because "learning machine" and "machine learning" have different implications. This logic might extend to trigrams or other n-grams. However, words that have differences in meaning when utilized in different contexts would be obscured through this

method. For example, the word python could represent a programming language, or a reptile.

The computer science community has identified a solution to these problems through a language model called BERT. BERT stands for Bidirectional Encoder Representations from Transformers. In 2018, when released, Devlin et al. (2018) achieved state-of-the-art performance on a number of NLP understanding tasks. Briefly, BERT embeddings are trained on the entirety of English language Wikipedia and 500 samples of text called the Book Corpus. The model is trained using two tasks: (1) masked language modeling, where 15 percent of tokens are masked and BERT was trained to predict them using the context, and (2) next sentence prediction, where BERT was asked to predict if a particular next sentence was probable given the first sentence. The purpose of these tasks is to output vectors for tokens that rely on context.

That is, when ingesting a job posting, each word (or subword) will be given a 768 dimensional vector based on the words around it. We could imagine that based on context, the vector for the word python when used as a programming language might be near other programming languages or words about debugging code, while the vector for the word python when used to describe reptiles might be near words for other snakes, like "boa," or words like "grass" or "slither."

In this paper, I currently utilize "pre-trained" BERT embeddings. That is, the vectors that are applied to each token (word or part of word) are based on English language Wikipedia and the Book Corpus. There are many reasons why the embeddings from these sources can convey similar meaning: both job postings and Wikipedia are written for relatively general audiences. Unlike patents or highly technical documents where the words represented may not even exist on Wikipedia, most words from job postings – which describe responsibilities, firm attributes, team composition, and educational requirements – represent similar concepts as they would on Wikipedia. There may be cases where words in job postings convey different meanings than on Wikipedia or in unpublished books. For example, the terms "preferred" and "desired" may convey similar meanings in job postings but different meanings in romance novels. However, for most words in job postings, I would conjecture that the meaning would be similar on Wikipedia and in novels. This is a testable prediction, and future iterations of this

work will develop pre-trained embeddings.

These pretrained BERT embeddings are the fundamental input of the NLP model predicting salaries, and therefore serve as the first layer. Because the BERT model has a length limit of 512 tokens, I only select the first 512 tokens of a job posting.[1]

More specifically, the model takes a job posting of 512 tokens as an input. A token in the BERT model is either a word, or a subword (part of the word), if the word is not sufficiently common. One estimate from another transformer model, GPT-3, suggests that, on average, 75 words consist of approximately 100 tokens. These 512 tokens are turned into a 512 by 768 dimensional matrix. This matrix is quite large, and the next layers in the model serve to reduce dimensionality. The model structure is displayed visually in Figure 4 and numerically in Table 1. First, a convolutional neural network summarizes each posting, by turning a single posting from 512 x 768 dimensions to 509 x 64 dimensions (taking four tokens at a time, conceptually condensing separate words into phrases). The next layer is a global max pooling layer, which takes the maximum value over dimensions, resulting in a 64 dimensional vector per posting. The next two layers flatten and normalize, concluding with an output layer that predicts the salary. Greater discussion on the layers of the model and the hyperparameters are described in the Appendix.

3.2 Model Evaluation

The model is currently trained on 855,477 postings from April 2019 to December 2019. The relevant evaluation metrics are based on the 214,281 postings that are "out-of-sample," i.e. not used in the training process. In data science terminology, this can be referred to as the "test" sample.

Table 2 compares models that do not use the full text of data to the fifth model, described above, that uses the full text of data.

The first column is a model with six-digit occupation fixed effects provided by BGT. [2] The

[1] In practice, this decision is not consequential: starting at a random point in the posting compared to starting at the beginning of the posting yields similar results. Longer postings seem to have more information about the application process and not actually about the job itself.

[2] Of the postings in the test sample, almost 95% of them are tagged with a six digit occupation label. The

coefficient of variation (R^2) on a simple regression containing occupation fixed effects for the sample is 0.590. This is notably much higher than an individual or household level regression on earnings (instead of at the posting level). However, there is still much left to be explained.

The next model, in Column (2), incorporates location. BGT postings are tagged with a best fit metropolitan statistical area (MSA). A fully interactive model, with separate fixed effects for each occupation by MSA, would capture the variation discussed in Deming and Kahn (2018), allowing for different local labor markets to have different skill requirements (and therefore, different wages) for different occupations. This model yields an R^2 of 69.5 percent, around 10 percentage points higher than a model with only occupation fixed effects.

Previous work has suggested that the skills articulated in job postings have predictive power for wages. A number of papers, including but not limited to Acemoglu et al. (2020) and Deming and Kahn (2018), utilize the skill data from BGT to characterize differences within occupation across firm or MSA.

Along these lines, Column (3) estimates a random forest regressor with the 28 BGT skill cluster families, listed in Appendix Table A1. A random forest regression is an ensemble model, where a number of decision trees are averaged to make a more accurate prediction. A longer discussion of random forests will be in the appendix. In this circumstances, 1000 different decision trees were fit on the training sample, and evaluated on the test sample.

The resulting increase in the explanatory power of salaries is modest – from 0.695 to 0.728. However, the skill cluster families provide meaningful information.

Column (4) uses the more granular Burning Glass Technologies skill clusters. There are 648 different skill clusters. For example, under the skill cluster family Information Technology, there are skill clusters of "Cybersecurity," "Technical Support," and "Java." Under the skill cluster families of Maintenance, Repair, and Installation, there are the skill clusters of "Vehicle Repair and Maintenance," "Hand Tools," and "Electrical and Mechanical Labor." The coefficient of variation continues to increase – to 0.765, a 3.7 percentage point increase.

Finally, Column (5) describes the natural language processing model, outlined in detail in

postings missing such a label are categorized as a separate category for the purpose of this analysis.

Section 2. The model performs substantially better – a coefficient of determination of 0.825. That is, the combination of words articulated in the job posting can explain 83 percent of the variation in metadata salaries, above and beyond the skills and occupation that come from the words.

The next row displays the same exercise for the Root Mean Square Error (RMSE). First, the RMSE falls only slightly between models (1) and (2), despite the sizable increase in the number of fixed effects. Moreover, the contrast between the fixed effects models and the NLP model is quite stark. The NLP model leads to a 39.3 percent decrease in the RMSE, compared to the occupation by MSA fixed effects specification.

4 Words and Phrases Associated with Salaries

Which words and phrases associated with higher and lower salaries? The NLP model described above is a salary prediction model, with hundreds of millions of parameters, making interpretability challenging. Even so, new techniques in the explainable AI literature can generate post hoc explanations of individual predictions made by such models (Agarwal et al., 2022). One such technique is called integrated gradients, which assigns an importance score to each input feature. Integrated gradients computes partial derivatives of the model output with respect to each input feature (Sundararajan et al., 2017). In the context of this paper, an input feature is a token (word or part of a word). Such features are not constrained to be the same within a posting – concretely, the word "help" might be positively related to salary when used in one context, while negatively related to salary in another context. This is consistent with the nuance employed when describing work in job postings.

As described in the data section, postings are capped at 512 tokens. This amounts to approximately 96 million tokens over the 214,281 postings in the test set. The added complication is that words can have different meanings - the word senior can be used to describe the experience level of a position, or used to describe the clientele of an assisted living facility. The task, therefore, of decomposing words into categories is quite complicated. I will start with a simple

descriptive analysis.

Before proceeding, it is important to reiterate that the output of the integrated gradients method is post hoc explanations, and these do not have a causal interpretation. However, the importance scores can provide insight into what characteristics might be valued in jobs.

4.1 Descriptive Analysis

LImiting the sample of tokens to those that contain letters, and aggregating tokens across all postings in the test set, produces 24,297 unique tokens. About half of these tokens are used less than one hundred times. The following figures use tokens that are mentioned at least one hundred times in the test set.

The weighted integrated gradient value for a given token over the test set is the sum of the integrated gradients divided by the count of that token. Returning to our example, the word "help" might be used positively and negatively within postings and across postings. The attributions are summed over all postings in the test set and divided by the frequency. This value has a straightforward interpretation: To use a concrete example from Figure 5, the inclusion of the word "intern" in a posting is, on average, associated with a 0.004 percent decrease in salary. This may be considered small, however, this is the effect of a *single* word.

The most positive and negative words are displayed in Figure 5. Of the fifteen most positive words, most are words related to job titles, such as "dentist," "practitioner," "professor," and "senior." There is one location related word in the most positive words too – "francisco," part of San Francisco, where the median household income is $119,136, close to double the median household income in the United States as of 2020.

The most negative words, in Figure 5b, are very different. These denote entry level or junior positions, such as "intern/internship," "assistant," and "entry." There are two words related to "leasing," and "cleaning," tasks that are considered low-skill. The list also includes titles such as "aide," "hostess," "bartender," "porter," "postdoctoral," and "scribe." The list also speaks to "childhood," likely referencing early childhood educators, who are one of the lowest-paid occupations.

Finally, the negative attributions list includes "Microsoft," likely referring to the skill of Microsoft Office Suite.[3]

In Figure 6, I characterize some of the tokens used into four different categories – locations, activities, technology skills, and titles. These are selected tokens but demonstrate reasonable patterns. For example, coastal cities and states are often associated with higher attributions, while midwestern and Appalachian locations are associated with lower attributions. The activities list shows that "manages," "evaluate," "repair," and "educate," are associated with positive attributions, while "clean," "plans," "assembles," and "patrolling," are associated with negative attributions. While the relationship between activities (also known as tasks) and wages are well-theorized (Acemoglu and Autor, 2011), and the task-based framework has become the dominant framework for thinking about the relationship between technology and wages, even leading research (e.g. Acemoglu and Restrepo, 2022) infers activity differences through changes in industrial composition of certain occupations. This paper is unique in directly measuring the activities being performed by workers and relating those activities to their wages.

The technology skills are almost all positive: "cloud," "sql," "java," and "python," for example, with the exception of the aforementioned, "microsoft," which has a strong negative attribution. This speaks to the hedonic and not causal nature of this model. Undoubtedly, a worker with additional skills will be more valuable, yet the selection into listing Microsoft Office Suite on a posting demonstrates something about the role.

To the extent that no such quantification of technology skills valuations has been developed, however, it may, in the future and with validation, be used to help make marginal human capital investment decisions.

Finally, Figure 6(d) looks at various title related tokens. "Retail," "driver," and "representative," are all associated with earnings decreases, while "nurse," "physician," and "administrator," are associated with earnings increases.

With these preliminary descriptive results in mind, I describe the decomposition exercise I

[3]While Microsoft also refers to the firm, the number of uses of Microsoft far outweights the postings for Microsoft firm postings.

hope to execute.

4.2 Decomposition

There are many determinants of earnings. Unfortunately, only few of these determinants are collected at a time in commonly used datasets: the publicly available Current Population Survey contains earnings, education, experience, occupation, and location, but leaves out firm names and skills that affect earnings. Despite being the gold standard for measurement of earnings, administrative datasets in the United States rarely have occupation.

Job postings have these characteristics and more. Above, I demonstrate that each token in each posting can be attributed a value. The next step is putting the 96 million tokens in the test set into categories: locations, job titles, experience levels, education levels, skills, activities, and firm names – all which have been shown to affect earnings.

Why the 96 million tokens instead of the 24,297 unique tokens in the test set? The Microsoft example is a cautionary one. Microsoft can represent a skill or a firm name. The context in which it is used determines which category it falls into. This is true of many other words - an internship can be listed in the requirements of a role, instead of pertaining to a job title, drastically affecting the interpretation.

My proposed solution to this gargantuan task is to hire a data labeling firm to label some fraction of the postings (5-10 percent) and then apply AutoML to the rest of the postings. This is the plan for early 2023.

5 Conclusion

This paper develops the first natural language processing model to predict the salary of job postings using the text. With new data on salaries from the metadata of job postings, the inputs and outputs are well-defined. This lends itself to the task of supervised machine learning, where the task is to derive the function that relates text to salaries. Because text in job postings is written in commonplace language, I use the technique called transfer learning – applying

knowledge gained from solving one problem to apply to this problem of salary prediction. In practice, this means that the first layer of my salary prediction model is pre-trained word embeddings from the BERT model, trained on English language Wikipedia and the Book Corpus.

My model substantially exceeds performance by any conventional baseline – a 39 percent decrease in RMSE and a 19 percent increase in R^2 compared to models with occupation by MSA fixed effects. This demonstrates that variation important for earnings can be found in the text of online job postings.

References

ACEMOGLU, D. AND D. AUTOR (2011): "Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings," Elsevier, vol. 4 of *Handbook of Labor Economics*, 1043 – 1171.

ACEMOGLU, D., D. AUTOR, J. HAZELL, AND P. RESTREPO (2020): "AI and Jobs: Evidence from Online Vacancies," Working Paper 28257, National Bureau of Economic Research.

ACEMOGLU, D. AND P. RESTREPO (2022): "Tasks, automation, and the rise in US wage inequality," *Econometrica*, 90, 1973–2016.

AGARWAL, C., E. SAXENA, S. KRISHNA, M. PAWELCZYK, N. JOHNSON, I. PURI, M. ZITNIK, AND H. LAKKARAJU (2022): "OpenXAI: Towards a Transparent Evaluation of Model Explanations," *arXiv preprint arXiv:2206.11104*.

AUTOR, D. H. AND M. J. HANDEL (2013): "Putting Tasks to the Test: Human Capital, Job Tasks, and Wages," *Journal of Labor Economics*, 31, S59–S96.

BLAIR, P. Q. AND D. J. DEMING (2020): "Structural Increases in Demand for Skill after the Great Recession," *AEA Papers and Proceedings*, 110, 362–65.

BOMMASANI, R., D. A. HUDSON, E. ADELI, R. ALTMAN, S. ARORA, S. VON ARX, M. S. BERNSTEIN, J. BOHG, A. BOSSELUT, E. BRUNSKILL, ET AL. (2021): "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*.

BONHOMME, S., K. HOLZHEU, T. LAMADON, E. MANRESA, M. MOGSTAD, AND B. SETZLER (2020): "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?" .

CARD, D., J. ROTHSTEIN, AND M. YI (2021): "Location, Location, Location," Working paper, U.S. Census Bureau Center for Economic Studies (CES).

DEMING, D. AND L. B. KAHN (2018): "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals," *Journal of Labor Economics*, 36, S337–S369.

DEMING, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market*," *The Quarterly Journal of Economics*, 132, 1593–1640.

DEVLIN, J., M.-W. CHANG, K. LEE, AND K. TOUTANOVA (2018): "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*.

DOMO (2018): "Data never sleeps 6: Domo," .

HECKMAN, J. J., L. J. LOCHNER, AND P. E. TODD (2006): "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," Elsevier, vol. 1 of *Handbook of the Economics of Education*, 307–458.

KUHN, P. AND H. MANSOUR (2014): "Is Internet Job Search Still Ineffective?" *The Economic Journal*, 124, 1213–1233.

MARINESCU, I. AND R. WOLTHOFF (2020): "Opening the Black Box of the Matching Function: The Power of Words," *Journal of Labor Economics*, 38, 535–568.

MINCER, J. (1974): *Schooling, experience, and earnings*, National Bureau of Economic Research.

SUNDARARAJAN, M., A. TALY, AND Q. YAN (2017): "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ed. by D. Precup and Y. W. Teh, PMLR, vol. 70 of *Proceedings of Machine Learning Research*, 3319–3328.

WEINBERGER, C. J. (2014): "The Increasing Complementarity between Cognitive and Social Skills," *The Review of Economics and Statistics*, 96, 849–861.

Figures and Tables

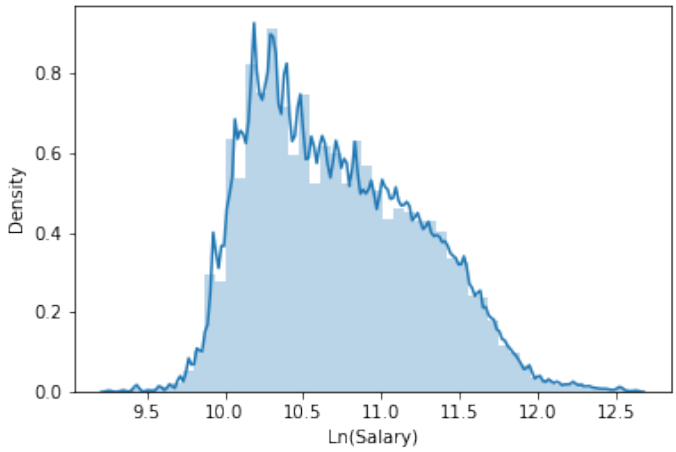


Figure 1: GHR Salary Distribution from April 2019 to September 2020

Notes: This figure describes the posted salary distribution of the 37,113,666 Greenwich.HR job postings with salary metadata posted between April 2019 and September 2020. The mean of the distribution is 52473.28.

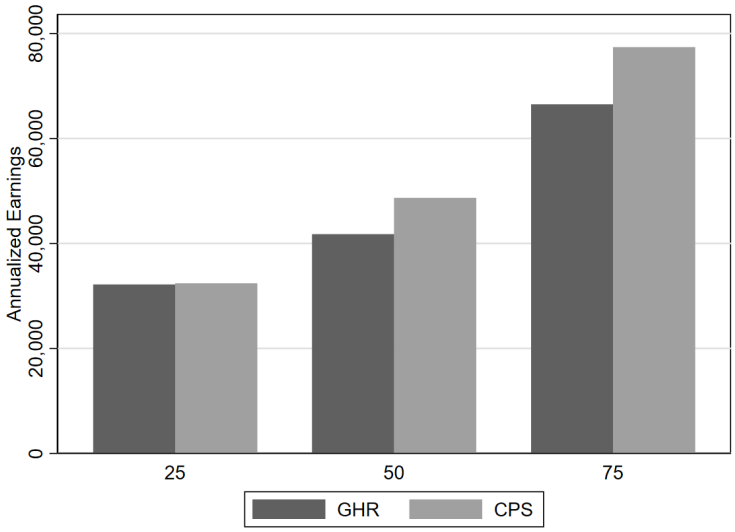


Figure 2: CPS - GHR Comparison

Notes: The Current Population Survey (CPS) values for quartiles of weekly earnings come from the Bureau of Labor Statistics' Usual Weekly Earnings of Wage and Salary Workers News Release Fourth Quarter 2019, available at `https://www.bls.gov/news.release/archives/wkyeng_01172020.htm`. CPS earnings are annualized by multiplying by 52. Data represent earnings before taxes and other deductions and include any overtime pay, commissions, or tips usually received (at the main job in the case of multiple jobholders). Greenwich.HR (GHR) salaries come from the full set of 37 million postings with salary metadata available.

Figure 3: Sample Job Postings in the Portland - Vancouver - Hillsboro Metropolitan Statistical Area

(a) Sales Floor Associate at Buy Buy Baby

'Sales Floor Associate\n\nBuy Buy BABY\n\n-\n\nBeaverton, OR 97005\n\nThe Sales Floor Associate oversees a Department within the store. In this role you will be a product, service and selling expert for your area while meeting sales and productivity goals.\nKey Responsibilities:\n\n* Models exceptional customer service by building relationships with store customers; makes appropriate recommendations based on customer needs; drives sales through suggestive selling, add-ons, and home deliveries\n* Meets with customers on a one-on-one basis to assist with determining personal needs and compiling merchandise preference list\n* Explains features of a broad array of merchandise to customers\n* Promptly and politely responds to customer inquiries and requests for support\n* Resolves customer issues using customer service skills, and escalates issues to more senior associates as necessary to ensure customer satisfaction\n* Organizes and straightens merchandise areas on the sales floor\n* Performs Registry Specialist tasks\n* Performs Sales Associate tasks\n* Knowledgeable of available technology and tools\n* Assists customers by offering a Baby order when merchandise is out of stock or not carried in the store\n* Performs additional duties as required including, but not limited to, stocking, freight processing, price changes, cart retrieval, break room and restroom housekeeping\n* Demonstrates commitment to the organization by maintaining regular, on site attendance, is reliable and follows through with responsibilities\n\nEducation/Experience:\n\n* High School diploma or equivalent\n* 2-4 years of retail experience desired\n\nsave this job a'

(b) Sales Associate at National Vision Inc.

'Sales Associate\n\nNational Vision, Inc.\n\n-\n\nVancouver, WA 98684\n\nPosition Description:\n\nAt National Vision, we believe everyone deserves to see their best to live their best. We help people by making quality eye care and eyewear more affordable and accessible.\n\nNational Vision, Inc. (NVI) is one of the largest optical retailers in the United States. We offer an innovative culture where training is a priority, hard work is praised, and career growth is a reality.\n\nWe are looking for a Sales Associate to join our growing team. The Sales Associate is responsible for selling, fitting and dispensing eyewear to customers.\n\nWhat would you do? The Specifics\n\n* Meet NVIs sales and company objectives.\n* Follow the Americas Best Code of Excellence to ensure customer satisfaction by creating a warm and welcoming environment for customers.\n* Assist with dispensing eyeglasses and contact lenses to customers, as permitted by state law.\n* Perform insertion and removal training of contact lenses to customers, as permitted by state law.\n* Educate clients on proper eyeglass and contact lens care.\n* Maintain accurate and organized patient records.\n* Assist Optometric Technician, Receptionist, and Contact Lens Technician when necessary.\n* Answer, screen, and forward incoming phone calls in accordance with NVI protocol.\n* Maintain visual merchandising according to Brand and Company Standards.\n\nPosition Requirements:\n\n* Previous retail experience preferred, but not required.\n* Maintain license, as required by state.\n* Strong selling skills, aimed at meeting both the stores and self-sales targets, by following company policies.\n* Strong customer service skills.\n* Able to give instruction in a clear and concise manner to customers.\n* Effective interpersonal skills.\n* Excellent organizational skills.\n* Detailed oriented.\n* Multitasking and time-management skills.\n* Ability to learn optical knowledge.\n* Professional attitude and appearance.\n* In some locations, bilingual abilities desired.\n\nWhat are the benefits?\n\nNational Vision offers a competitive benefits package including Health and Dental Insurance, 401k with company match, Flex Spending Account, Short Term and Long Term Disability Insurance, Life Insurance, Paid Personal Time Off, and much more. Please see our website at www.nationalvision.com to learn more.\n\nsave this job a'

Notes: The job text of two sample postings in raw form from Burning Glass Technologies.

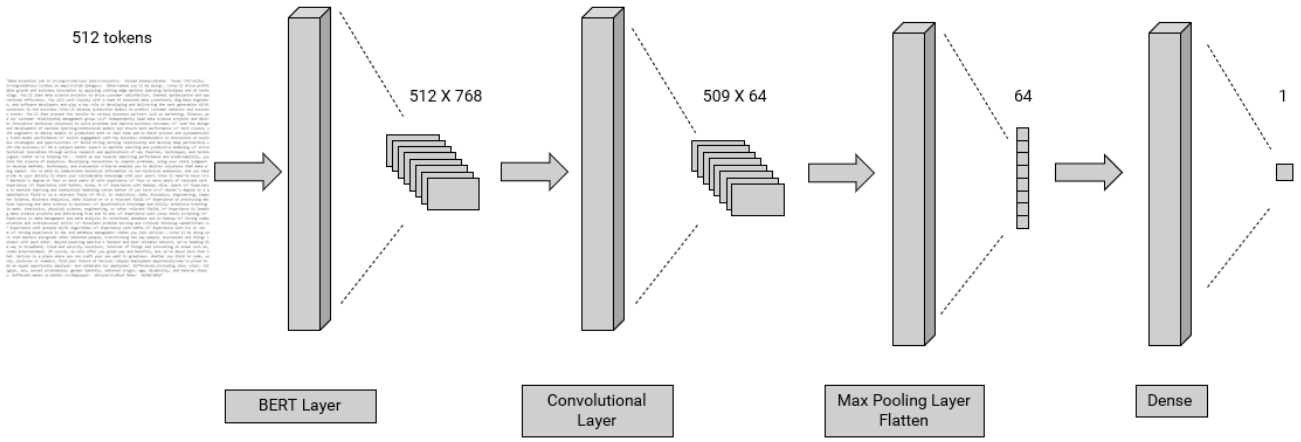
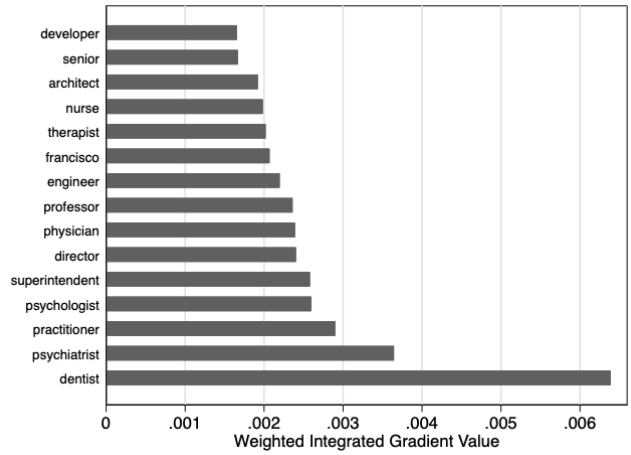


Figure 4: Model Structure

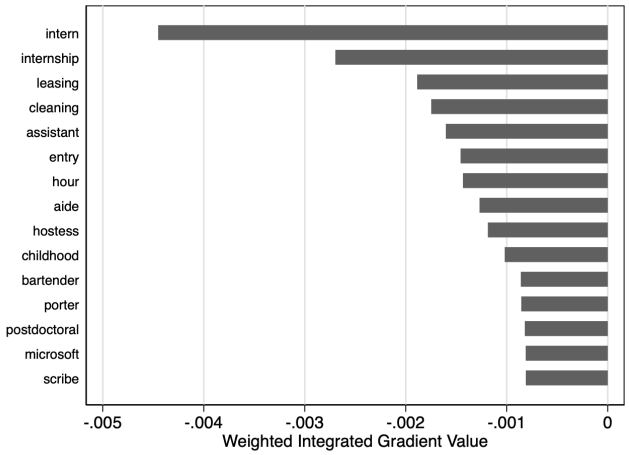
Notes: This figure displays the model structure. A job posting of 512 tokens is the input. BERT embeddings take these 512 dimensions and assign each a 768 dimensional vector depending on context. The next layer is a one dimensional convolutional layer. It is ultimately identifying n-grams that are predictive. The resulting matrix is 509 x 64. The next layer is a global max pooling layer, which captures the most relevant features from a sentence. This layer is flattened and turned into a 64 dimensional vector, which eventually predicts one dimensional ln(salary). The parameters are also laid out numerically in Table 1.

Figure 5: Words with the Most Positive and Negative Attributions

(a) Positive Attributions



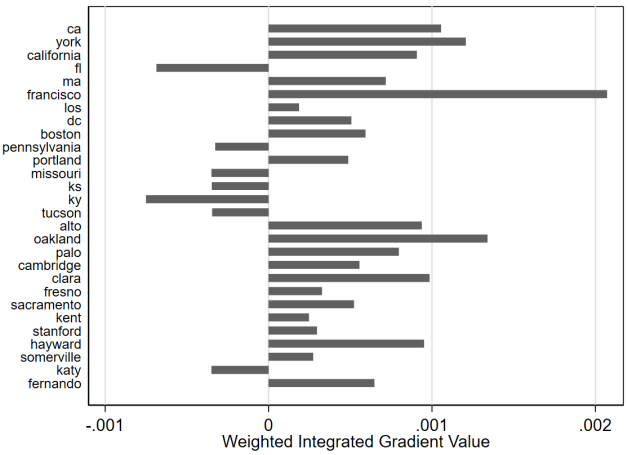
(b) Negative Attributions



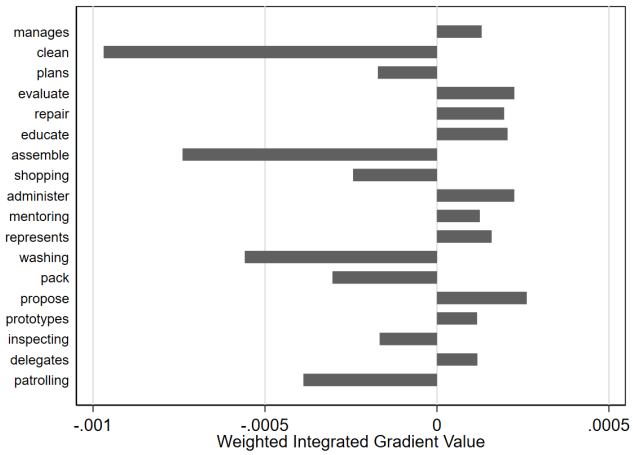
Notes:

Figure 6: Select Tokens in Postings and Their Integrated Gradients

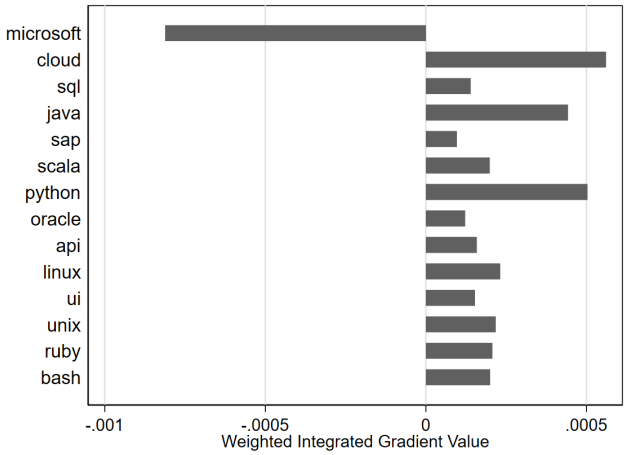
(a) Locations



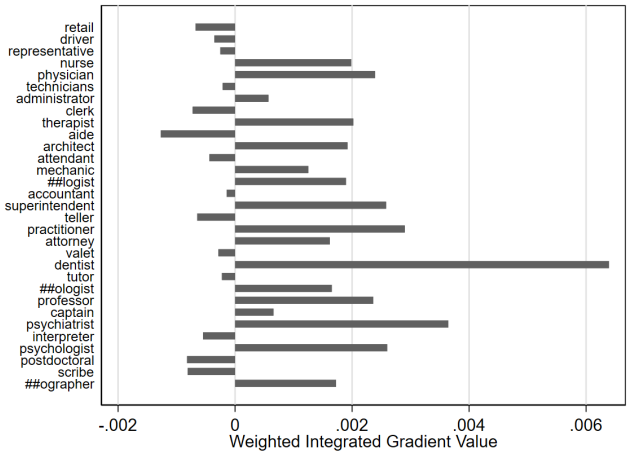
(b) Activities



(c) Technology Skills



(d) Titles



Notes:

Table 1: Model Architecture

Layer Type	Dimensions	Number of Parameters
Input Layer	(X, 512)	
BERT Layer	(X, 512, 768)	109482240
Convolutional Layer	(X, 509, 64)	196672
Global Max Pooling Layer	(X, 64)	
Flatten	(X,64)	
Batch Normalization	(X, 64)	256
Dense Layer	(X, 64)	4160
Dense Layer	(X, 1)	65
Total params: 109,683,393		
Trainable params: 201,025		
Non-trainable params: 109,482,368		

Notes: This table describes the architecture of the natural language processing model used to predict salaries. In this table, X denotes the number of postings fed into the model. The input is 512 tokens of a job postings from October 2019. These 512 tokens are fed into a BERT embedding layer, where each token is given a 768 dimensional vector that is context dependent. At this point, each posting has 512 x 768 dimensions – likely too many inputs to a single salary value, so the next layers are focused on condensing dimensionality. The first step is a convolutional layer, which takes 512 x 768 dimensions, and reduces it to 509 x 64. The next layer, a global max pooling layer, takes the maximum values from this 509 x 64 matrix, which can be perceived as the most salient features, and condenses it to just 64 dimensions. The following two layers flatten and normalize these layers. Eventually, these 64 dimensions are condensed to a single dimension - the natural log of salary.

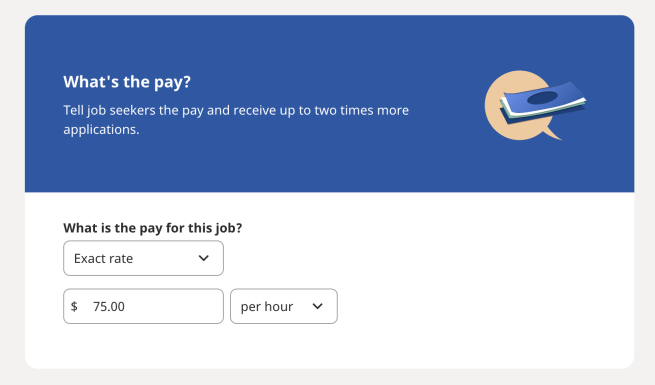
Table 2: Out-of-Sample Performance Metrics

	(1) Occupation FEs	(2) Occupation x MSA FEs	(3) Occupation, MSAs, & Skill Cluster Families	(4) Occupation, MSAs, & Skill Clusters	(5) NLP Model
R^2	0.590	0.695	0.728	0.765	0.825
RMSE	0.330	0.317	0.269	0.249	0.200
Occupations	785	785	785	785	
Locations	-	807	807	807	
Skill Categories			28	648	

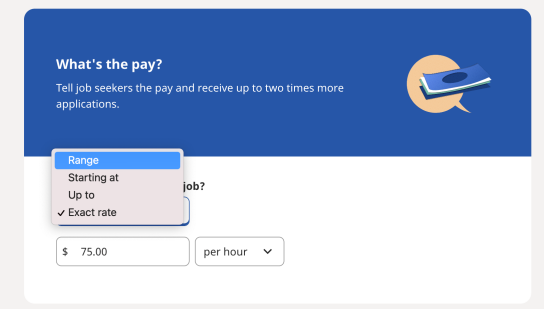
Notes. This table summarizes the performance of the natural language processing model, in Column (5), to a number of relevant baselines. Relevant metrics are R^2 (coefficient of variation) and Root Mean Square Error (RMSE). The entire test set (214,281 observations) is used in every model. The outcome is ln(salary). Column (1) includes six digit Standard Occupation Classification (SOC) fixed effects. Column (2) interacts these occupation fixed effects with MSA fixed effects. Column (3) estimates a random forest regressor model with Burning Glass Technologies Skill Cluster Family categories. These are listed in the Appendix. Column (4) replaces the skill cluster family categories with Skill Clusters. There are 648 skill clusters. Column (5) is the model described extensively in Section 2.

Figure A1: Screenshots of Job Board User Interfaces for Recruiters To Input Salaries

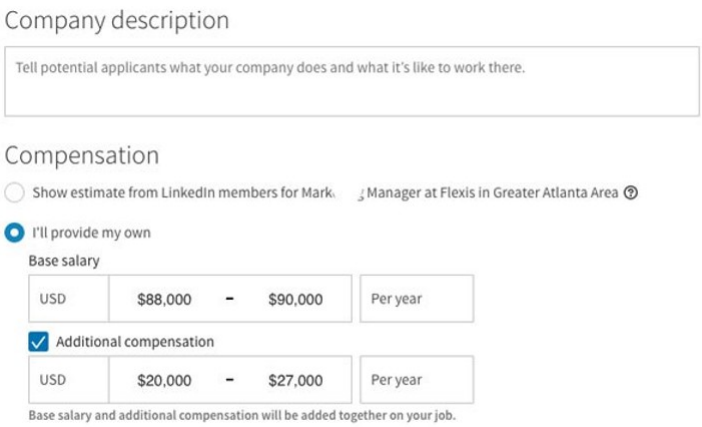
(a) Indeed Posting Screen for Recruiters



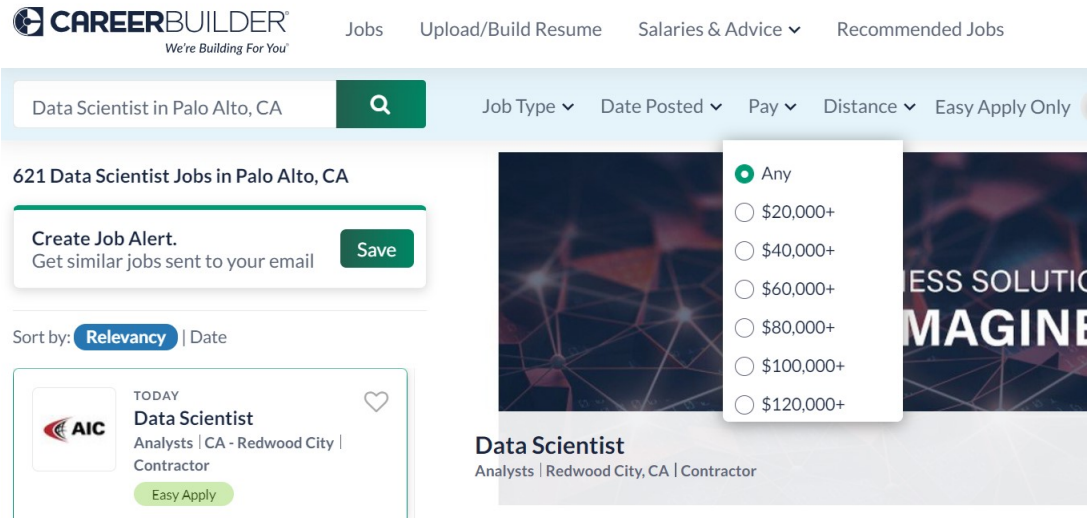
(b) Indeed Options for Recruiters



(c) LinkedIn Compensation Screen for Recruiters



(d) Career Builder Search Portal for Applicants



Notes: This figure demonstrates recruiter side of job posting platforms, which provide the opportunity for recruiters to input salaries. In Panel A, a recruiter is asked the pay for the job. They are incentivized by the statement, "Tell job seekers the pay and receive up to two times more applications." In Panel B, options are displayed. A recruiter can input a range, starting at, up to, or an exact rate. In Panel C, this is the screen on the popular site, LinkedIn. Recruiters are even asked for base salary and additional compensation in separate fields. Finally, in Panel D, you can see the applicant side on another platform, CareerBuilder. The search tool allows applicants to search above a certain pay threshold.

Table A1: Skill Cluster Families

Administration	Human Resources
Agriculture, Horticulture, and the Outdoors	Industry Knowledge
Analysis	Information Technology
Architecture and Construction	Legal
Business	Maintenance, Repair, and Installation
Customer and Client Support	Manufacturing and Production
Design	Marketing and Public Relations
Economics, Policy, and Social Studies	Media and Writing
Education and Training	Personal Care and Services
Energy and Utilities	Public Safety and National Security
Engineering	Religion
Environment	Sales
Finance	Science and Research
Health Care	Supply Chain and Logistics

List of Skill Cluster Families used in Column (3) of Table 2.