CS230

Studying the Relationship between Social Position and the Meaning of Work using Siamese BERT Networks and Generalized K-Means Clustering

Sheridan Stewart
Department of Sociology
Stanford University
sastew@stanford.edu

Abstract

It is often taken for granted that work means the same thing to everyone. However, at a time of near-unprecedented political polarization and in the midst of a recession in which tens of millions of people have lost their jobs, it is imperative that we develop a greater understanding of not only what work means to different people, but also how the meaning of work systematically relates to workers' positions within society and economic risks they face. To uncover the distribution of expectations about work across the occupational structure, I analyze workers' review of their employers, first developing rich document embeddings and then clustering occupations. First, I conduct an experiment to contrast the ability of a multilayer, bidirectional LSTM network and four variations of Siamese BERT networks to classify reviews into one of 49 occupations. I find that using my data to fine-tune a Sentence-BERT network pre-trained for NLI and STS-B performs best. Next, I use this model to encode the reviews before applying a novel clustering algorithm that preserves information about workers' stances toward their experiences as evidence the one- to five-star ratings they provide. I am thus able to test whether there is evidence of consensus about "bad jobs" while exploring the distribution of expectations about work over the occupational structure.

1 Introduction

In the midst of an economic recession that has seen tens of millions of workers lose their jobs while others now work remotely or risk contracting COVID-19, little attention has been paid to how the meaning of work can fundamentally shape how individuals experience transitions of this kind [1, 2, 3, 4]. Even less attention has been paid to the potential of systematic relationships between workers' positions within society and what work means to them.

Occupations have been identified as one grouping of relatively culturally homogeneous individuals, and they relate directly to members' social positions [5]. Although researchers have analyzed how occupations are perceived from the outside [6, 7], they have neglected the what work means of to *members* of these occupations.

To uncover variation in the meaning of work across the occupation structure, I use the text of Glassdoor reviews (i.e., reviews of employers submitted by employees) and their occupational labels as inputs to a series of classifiers in order to identify the best model for embedding the reviews and clustering the occupations. Specifically, I compare a two-layer Bi-LSTM network to four variations on Sentence-BERT models (SBERT) [8], an approach to Siamese BERT networks. Rather than treat

CS230: Deep Learning, Fall 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

reviews as if they are equivalent, I preserve information about the explicit *stance* that workers take toward their employers provided in their employer rating (out of five stars). To do this, I utilize a novel generalization of the k-means algorithm to cluster occupations based on stratified (1-star to 5-star) subsets of the data.

2 Related work

Meaning of Work. A wealth of research exists on how workers of different types–skilled vs. unskilled labor, blue- vs. white-collar, etc.–are stereotyped by observers (e.g., "computer programmer" as a masculine occupation) [6, 7]. However, in a sprawling, fragmented literature on the meaning of work to workers [2], there is a relative lack of attention to variation in the meaning of work. This can be seen in claims about relative consensus about "bad jobs" [3], but also in research geared to improving productivity. Moreover, this literature rarely links the meaning of work to social position. In this project, I am guided by discussions of the relationship between social position and the meaning of work in limited recent research [4] and in past work that posited three broad orientations toward work: as a job (making ends meet), as a career (increasing prestige and influence), and as a calling (a way to find fulfillment and oneself) [1].

Linking Meaning to Social Position. Increasingly, inequality researchers have moved away from broad categories based on social class and toward studying institutions such as occupations as "micro-classes": as a result of their social origins, individuals face different constraints (e.g., educational opportunities, networking), have different interests, and so on. Consequently, they are effectively sorted into–and select into–different occupations [5]. As they enter their occupations and on the job, workers are further socialized by the types of work they do, the people with whom they have opportunities to interact, and so forth. In short, there are numerous ways in which work is an important driver of changing attitudes, tastes, and beliefs. Occupations are therefore an ideal unit of analysis for studying how variation in the meaning of work relates to social position.

Research on other focal experiences points to the merits of linking social position to meaning, for example research on dining out at restaurants that emphasized differences in the narratives that exist among positive and negative experiences, as reflected in the text and ratings of Yelp reviews [9].

3 Dataset and Features

My corpus consists of approximately 2.7 million reviews of companies submitted to Glassdoor.[1] Each review contains text about "pros" and "cons" of working for the company as well as an overall rating from 1-5 stars. These data were provided by Glassdoor for my dissertation, of which this project will become a part. Reviews were coded into 158 usable occupational categories by Glassdoor;this is tremendously helpful, though I caution that in a future iteration of this work, workers' self-described job titles will be mapped to an existing taxonomy of occupations more widely used by inequality researchers. Results here–which depend not only on potentially different labels, but also on different data–may not reflect future results, whether or not these are encouraging.

The "pros" and "cons" sections of the reviews reveal which of workers' expectations are met ("pros") and violated ("cons"), shedding light on what work means to the worker. Further, workers often write about positive or negative things in the opposite section (e.g., "Despite the great work-life balance..." in the "cons" section). As a result, I concatenated these sections to form a single document for each review. I kept each document with at least 10 words; to fit my chosen maximum sequence length (128) for training several models in my experiment, I used the first and last 64 words of any document $>$ 128 words. Additionally, I conducted only minimal preprocessing, such as expanding contractions (e.g., "didn't" to "did not"). For the LSTM baseline (described below), I additionally lemmatized the text before feeding it into an embedding layer with GloVe embeddings [10].[2]

Given the considerations above, I used a stratified sampling strategy based on a minimum number of reviews for each rating (e.g., 1-star) for each occupation. Different minima resulted in trade-offs between the number of occupations with enough reviews and the overall corpus size; these trade-offs

[1] https://www.glassdoor.com

[2] Specifically, I used the 300-dimensional embeddings with a 1.9 million-token vocabulary that was pretrained on Common Crawl data.

Review	Occupation
Flexible working hours and great working environment. Provides employees with benefits comparable to industry standards [...] Employees can relax by playing ping pong, pool table, bowling alleys, fitness room...Fun parties..	Software Engineer
COMPANY is large company - many opportunities to move laterally and up! COMPANY is very good with work-life balance and respects it. They encourage further education and offer training. [...] Buildings and facilities are outdated (furniture, layout), however getting better with updating technology.	Engineer
Very little overtime. Hot. Very hard work. 12 hour shifts 3 days one week and 4 days next. Very little sitting.	Unskilled

Table 1: Example reviews illustrating different work experiences and priorities.

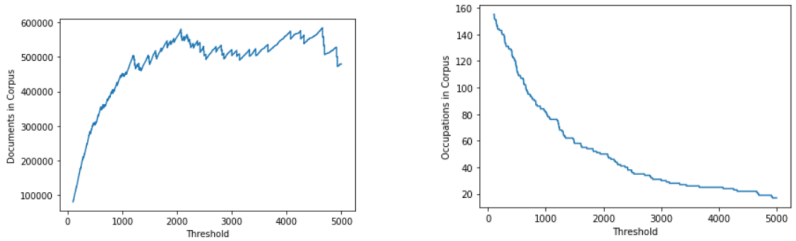

Figure 1: Left: Increase in corpus size (number of reviews) as the minimum number of reviews per occupation increases. Right: Decrease in the number of occupations in corpus as this threshold increases.

are depicted in Figure 1. I selected a threshold of 2,000 for each rating for each occupation, resulting in a minimum of 10,000 reviews per occupation (2,000 per rating). After excluding the "other" occupation category, there remained sample of 49 occupations (listed in Table A1 in the appendix). I divided the reviews for these occupations into a training set (8,000 reviews per occupation, 392,000 total) and a development set (2,000 reviews per occupation, 98,000 total), and then selected a test set from the remaining documents for these occupations (1,000 reviews per occupation, 49,000 total).

Table 1 presents three example reviews with their occupational labels as they appear in the data set. Notably, the second review focuses much more on work as a career (one possible work orientation). In the third review, we see dismay about a lack of overtime–an indication of a "work as a job" orientation, rather than a perception of work as a ladder to climb or a source of personal fulfillment.

4 Methods

Prior to clustering occupations, I conducted an experiment to identify which of several models could best distinguish between reviews from the 49 distinct classes. The aim was to use the model best able to distinguish the occupations in order to encode representations of the reviews that capture what is unique about each occupation; subsequently clustering them will, then, group occupations based on the relatively "unique" traits they have in common. This represents a more rigorous test of the idea that the meaning of work transcends occupational cultures.

Baseline model. Because the data are proprietary, I am unable to share them with others in order to get an estimate of human-level performance. Rather than compare only a set of SBERT models, I use a two-layer Bi-LSTM network as a baseline (Model 1. There is a long history of using RNNs, including those in the LSTM family, for text classification, where they remain useful [11].

Because the SBERT models would have an advantage due to transfer learning, I made this comparison more challenging by using a GloVe embedding layer (described above). The network diagram on the left-hand side of Figure 2 depicts the full architecture of this model. A LayerNorm layer [12] followed the embedding layer, and this was followed by two Bi-LSTMs with 50% dropout using Tanh. The network then progressed to a flatten layer, a Tanh layer, 20% dropout, a second Tanh layer,

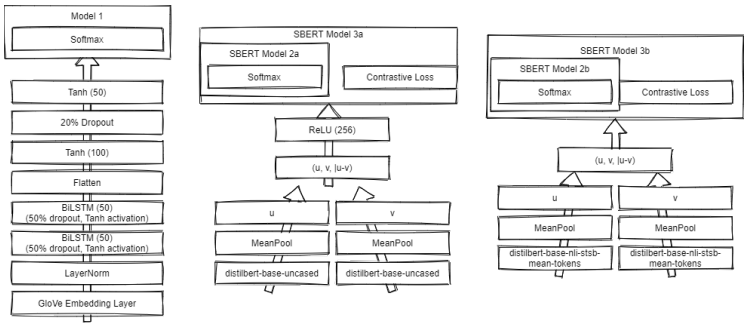

Figure 2: Left: Two-layer Bi-LSTM (Model 1). Center: Training SBERT models with a softmax loss (2a) and both softmax and contrastive loss (2b). Right: Fine-tuning pretrained SBERT models with a softmax loss (3a) and both a softmax loss and contrastive loss (3b).

and finally a softmax layer that output a prediction from among one of the 49 classes. This structure was chosen after much trial and error, during which it was difficult to get an LSTM-based network to perform better than chance. During this process, I iterated through combinations of randomly sampled learning rates and four optimizers (Adam, RMSProp, Adagrad, and SGD). I found that Adagrad with a learning rate of 0.1 performed best for the dev set. This network was trained on the full training set for 50 epochs using a categorical cross-entropy loss and was evaluated on the test set only once (Table 2).

Sentence-BERT. The four main competitors in my experiment were variations on Sentence-BERT (SBERT) [8], a Siamese network structure. Specifically, I tested two SBERT models trained "from scratch" on my data set using `distilbert-base-uncased` [13] and two models pretrained for NLI and STS-B (`distilbert-base-nli-stsb-mean-tokens`) that I fine-tuned on my data. The models I trained "from scratch" used the conventional SBERT Siamese structure, with separate documents fed into the network, pooled, and concatenated (the first vector, the second vector, and the absolute difference) before being passed into a ReLU layer to reduce dimensionality.

Each SBERT model used a softmax layer with a categorical cross-entropy loss. However, one model trained from scratch (2b) and one fine-tuned, pre-trained model (3b) were trained iteratively with both a categorical cross-entropy loss and a contrastive loss [14, 15]. Each SBERT model was optimized using Adam with decoupled weight decay optimizer [16].

Stratified k-means. Because I wished to cluster occupations in a way that preserved information about the stances reviewers took toward their employers, my approach to clustering involves stratifying the corpus into a subset for each rating from one to five stars. I then use a novel clustering algorithm to generalize k-means to this multi-layer network. For each value of k, I apply the k-means algorithm to a single subset of the data. I then assign occupations to clusters, apply those assignments to the subsets of data for each of the other four ratings, calculating each subset-specific within-cluster sum of squares (inertia), and then average these. This score is stored for each (k, starting subset) pair. I then select the best value of k and the best starting subset. This approach is depicted in Equation 1.

$$(1) \quad argmin_k \frac{1}{S} \sum_{s=1}^{S} \sum_{i \in c_k} ||x_s^{(i)} - m_{ks}||^2$$

5 Results and Discussion

Table 2 presents the results of the experiment. The two-layer Bi-LSTM (Model 1) is competitive with the SBERT models trained using both a softmax and constrative loss (Models 3a and 3b), though the SBERT models trained only with a softmax loss (2a and 2b) perform better on all metrics except for macro-averaged precision. Although the two-layer Bi-LSTM has the highest macro-averaged precision, the SBERT model fine-tuned from the pre-trained model trained for NLI and STS-B performed best on all other metrics. This model's prediction were more accurate than chance by a factor of 14.9 with an accuracy of 0.305.

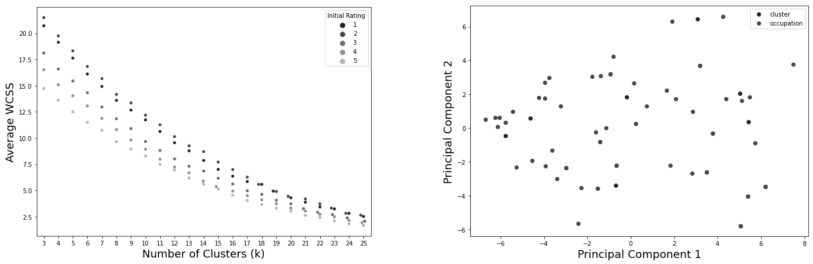

Figure 3: Left: Average within-cluster sum of squares (intertia) from generalization of k-means algorithm to corpus stratified by rating. Right: Clustering solution plotted using PCA with two components. Cluster centroids in black and occupation centroids in gray.

It is noteworthy that the SBERT models struggled to make predictions for some of the classes, whereas the two-layer Bi-LSTM did not. This can be seen in Table A1 in the Appendix, which shows the classification reports for the two-layer Bi-LSTM and the best-performing SBERT model.

Using the best of these models, I encoded the review from the train and development set in order to test my clustering algorithm. Figure 3 depicts these results. The left-hand panel shows the decline in the averaged within-cluster sum of squares (inertia) as the number of clusters increases. This near-exponential decay pattern was also apparent in the k-means results using the full data set (i.e., not stratified by rating). Although these results are not promising, I include in the right-hand panel a visualization of the clustering solution: black dots represent cluster centroids, while gray dots represent each occupation (as the average of its documents).

Model	**Accuracy**	$>$ chance	**Precision**	**Recall**	**F1 Score**
Softmax Loss					
1. Two-layer Bi-LSTM	0.187	9.2x	**0.197**	0.187	0.171
2a. SBERT (New)	0.274	13.4x	0.169	0.273	0.197
2b. SBERT (Fine-tuned)	**0.305**	**14.9x**	0.180	**0.306**	**0.219**
Softmax + Contrastive Loss					
3a. SBERT (New)	0.187	9.2x	0.088	0.187	0.096
3b. SBERT (Fine-tuned)	0.191	9.4x	0.066	0.191	0.093

Table 2: Accuracy, improvement over chance, and macro-averaged precision, recall, and f1 scores. Models 2a and 3a were trained on the Glassdoor corpus using distilbert-base-uncased. Models 2b and 3b were fine-tuned on the Glassdoor corpus using distilbert-base-nli-stsb-mean-tokens.

6 Conclusions and Future Work

In this paper, I have presented the results of an experiment contrasting five classifiers on the basis of their performance on a difficult mult-class classification problem. Because I am unable to share this proprietary data set, I am unable to estimate human-level performance; however, using a competitive two-layer Bi-LSTM as a baseline, I have provided further evidence of the merits of the Sentence-BERT approach [8] and DistilBERT [13], as well as the more general promise of both transfer learning document embeddings for studying meaning and culture in areas of great interest to scholars in various fields, including the social sciences.

The results of my clustering algorithm are not promising, though a comparison to running k-means on the full corpus shows the same exponential decay pattern in the weighted sum of squares (inertia) as the number of clusters increases. This may suggest that either k-means in unsuited for this problem, or that this distinguishing occupations on the basis of these reviews is simply a difficult task. My experimental results provide a starting point for further refinement of models that may better classify occupations. Future will also involve coding the job titles that reviewers themselves provided to match a taxonomy of occupations that is commonly used in research on inequality and stratification. Collecting more data, using labels that have long been vetted by other researchers, and improving

classifiers may lead to an improved ability to use clustering algorithms such as k-means. This future work will undoubtedly provide much-needed insight into the relationship between what work means to workers and their positions within society.

7 Contributions

Sheridan Stewart conducted this work on his own using a variety of Python libraries for scientific research and natural language processing [17, 18, 19, 20, 21].

References

[1] Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., Tipton, S. M. (1996). *Habits of the heart*. University of California Press.

[2] Rosso, B. D., Dekas, K. H., Wrzesniewski, A. (2010). On the meaning of work: A theoretical integration and review. *Research in Organizational Behavior*, *30*, 91-127.

[3] Kalleberg, A. L. (2011). *Good jobs, bad jobs: The rise of polarized and precarious employment systems in the United States, 1970s to 2000s*. Russell Sage Foundation.

[4] Sharone, O. (2013). *Flawed system/flawed self: Job searching and unemployment experiences*. University of Chicago Press.

[5] Weeden, K. A., Grusky, D. B. (2005). The case for a new class map. *American Journal of Sociology*, *111*(1), 141-212.

[6] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. 3*0th Conference on Neural Information Processing Systems (NIPS 2016)*.

[7] Garg, N., Schiebinger, L, Jurafsky, D., & James Zou (2018).Word embeddings quantify 100 years of genderand ethnic stereotypes. *PNAS*, *115*(16), E3635–E3644.

[8] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

[9] Jurafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, *19*, 4-7.

[10] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

[11] Wang, C., Singh, O., Tang, Z., & Dai, H. (2017). Using a recurrent neural network model for classification of tweets conveyed influenza-related information. *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017*, 33–38.

[12] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv:1607.06450v1.

[13] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS'19*.

[14] Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality deduction by learning an invariant mapping. *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

[15] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. arXiv:2004.11362v3.

[16] Loshchilov, I., & Hutter, F. (2019). Fixing weight decay regularization in Adam. *ICLR*. arXiv:1711.05101v3

[17] Pedregosa et al. (2011). Scikit-learn: Machine learning in Python. *JMLR 12*, 2825-2830.

[18] Abadi, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/

[19] Paszke, A., et al. (2019). PyTorch: An imperative style, high-Performance deep learning library. *Advances in Neural Information Processing Systems 32*, 8024-8035.

[20] Chollet, F., et al., (2015). Keras. https://keras.io

[21] Wolf, T., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45.

Appendix

	Two-layer Bi-LSTM			SBERT (2a)		
Occupation	F1 Score	Precision	Recall	F1 -Score	Precision	Recall
Academic Counselor	0.243	0.284	0.212	0.000	0.000	0.000
Account Executive	0.221	0.201	0.246	0.323	0.220	0.612
Accounting	0.109	0.074	0.202	0.212	0.136	0.480
Administrative	0.071	0.064	0.078	0.000	0.000	0.000
Analyst	0.007	0.049	0.004	0.000	0.000	0.000
Beauty	0.333	0.399	0.286	0.601	0.489	0.778
Branch Manager	0.004	0.182	0.002	0.000	0.000	0.000
Business Analyst	0.088	0.064	0.138	0.114	0.088	0.160
Claims	0.194	0.127	0.403	0.398	0.282	0.676
Client Services	0.008	0.053	0.004	0.000	0.000	0.000
Corporate Account Manager	0.026	0.088	0.015	0.000	0.000	0.000
Customer Service	0.099	0.092	0.107	0.000	0.000	0.000
Designer	0.238	0.436	0.164	0.453	0.370	0.584
Driver	0.391	0.408	0.375	0.633	0.559	0.730
Editor	0.302	0.457	0.225	0.574	0.539	0.614
Engineer	0.118	0.170	0.091	0.000	0.000	0.000
Field Sales Manager	0.101	0.099	0.103	0.000	0.000	0.000
Finance Specialist	0.140	0.210	0.105	0.381	0.364	0.400
Food Services	0.151	0.204	0.120	0.370	0.277	0.560
Front Desk	0.138	0.141	0.135	0.000	0.000	0.000
HR Specialist	0.060	0.058	0.062	0.000	0.000	0.000
IT	0.155	0.151	0.158	0.184	0.115	0.458
Management Consulting	0.234	0.236	0.231	0.396	0.322	0.514
Marketing Manager	0.103	0.088	0.124	0.000	0.000	0.000
Medical Technician	0.132	0.340	0.082	0.000	0.000	0.000
Merchandiser	0.161	0.263	0.116	0.279	0.212	0.409
Nursing	0.402	0.377	0.430	0.439	0.291	0.898
Operations	0.013	0.084	0.007	0.000	0.000	0.000
Patient Care Technician	0.259	0.196	0.381	0.000	0.000	0.000
Police & Security Officers	0.278	0.228	0.355	0.547	0.445	0.708
Program Coordinator	0.058	0.091	0.043	0.000	0.000	0.000
Project Manager	0.033	0.118	0.019	0.000	0.000	0.000
Quality Assurance	0.033	0.097	0.020	0.000	0.000	0.000
Recruiter	0.192	0.134	0.341	0.458	0.383	0.570
Researcher	0.302	0.467	0.223	0.449	0.352	0.619
Retail Representative	0.116	0.129	0.105	0.000	0.000	0.000
Sales Representative	0.095	0.162	0.067	0.000	0.000	0.000
Server	0.392	0.290	**0.608**	**0.690**	**0.703**	0.678
Skilled Labor	0.159	0.121	0.231	0.000	0.000	0.000
Software Engineer	0.218	0.224	0.212	0.389	0.300	0.554
Stock Clerk	0.214	0.156	0.342	0.378	0.264	0.660
Store Manager	0.283	0.249	0.327	0.444	0.371	0.552
Systems Administrator	0.043	0.123	0.026	0.000	0.000	0.000
Systems Technician	0.035	0.138	0.020	0.000	0.000	0.000
Teacher	**0.569**	**0.550**	0.590	0.667	0.529	**0.900**
Technical Support	0.115	0.195	0.082	0.000	0.000	0.000
Teller	0.385	0.291	0.571	0.587	0.450	0.845
Underwriter	0.196	0.201	0.192	0.501	0.570	0.446
Unskilled Labor	0.144	0.112	0.201	0.278	0.181	0.600

Table A1. Relative Performance of two-layer Bi-LSTM (Model 1) and pretrained SBERT fine-tuned on my data set with a softmax loss (Model 2a) based on F1 scores, precision, and recall. Best scores for each column bolded.