280 Characters to Employment: Using Twitter to Quantify Job Vacancies

Boris Sobol[1], **Manuel Tonneau**[2,3,4], **Samuel Fraiberger**[3,4,5], **Do Lee**[4], **and Nir Grinberg**[1]

[1] Ben-Gurion University of the Negev,
[2] University of Oxford,
[3] The World Bank,
[4] New York University,
[5] Massachusetts Institute of Technology
sobol@post.bgu.ac.il, manuel.tonneau@oii.ox.ac.uk, dql204@nyu.edu, sfraiberger@worldbank.org, nirgrn@bgu.ac.il

Abstract

Accurate assessment of workforce needs is critical for designing well-informed economic policy and improving market efficiency. While surveys are the gold standard for estimating when and where workers are needed, they also have important limitations, most notably their substantial costs, dependence on existing and extensive surveying infrastructure, and limited temporal, geographical, and sectorial resolution. Here, we investigate the potential of social media to provide a complementary signal for estimating labor market demand. We introduce a novel statistical approach for extracting information about the location and occupation advertised in job vacancies posted on Twitter. We then construct an aggregate index of labor market demand by occupational class in every major U.S. city from 2015 to 2022, which we evaluate against two sources of official statistics and an index from a large aggregator of online job postings. We find that the newly constructed index is strongly correlated with official statistics and, in some cases, advantageous compared to statistics from job aggregators. Moreover, we demonstrate that our index can robustly improve the prediction of official statistics across occupations and states.

Introduction

Timely information about labor demand is essential for policy-makers to assess which sectors require support during an economic slowdown or can benefit from educational training programs to reduce skill mismatch. However, estimates of such demand are traditionally based on surveys, which are costly to produce and often lack timeliness, particularly in developing countries (Devarajan 2013). This low frequency also implies a limited ability of surveys to capture rapid changes in the labor market such as those occurring during economic downturns or major technological shifts.

The growing prevalence of online job ads has made digital data a valuable complement to official statistics for timely information about labor market dynamics (Carnevale, Jayasundera, and Repnikov 2014). For instance, Burning Glass Technologies (BGT), a job posting aggregator that is estimated to have captured roughly 35% of all U.S. job postings between 2012 and 2018 (Cammeraat and Squicciarini 2021), is increasingly used in research on labor market demand (Deming and Kahn 2018). However, data from such platforms is not publicly available, and, more importantly, it provides a picture of the labor market that is not necessarily complete or representative. Indeed, such indices are typically skewed towards large companies, formal employment, and developed countries (Zhu, Fritzler, and Orlowski 2018).

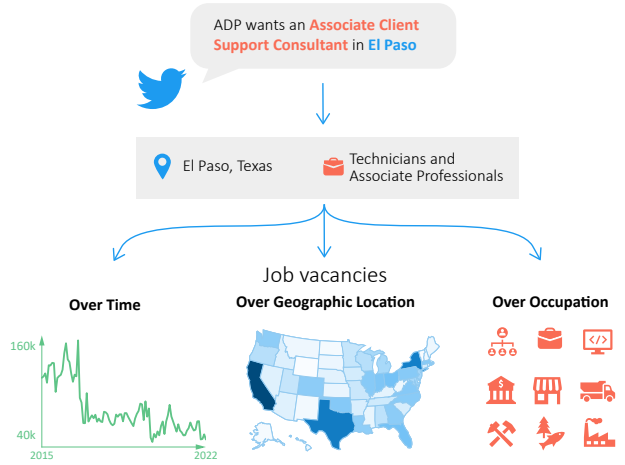

Figure 1: The pipeline for extracting job opening characteristics from tweets and constructing an index of job openings. Extraction was conducted by developing a fine-tuned BERT-based NER model to identify job occupations in addition to locations and organization entities. An index was created by aggregating tweets over time in different locations and occupations.

In this context, non-specialized social media platforms such as Twitter or Facebook represent another alternative data source to gain insights into labor market demand and complement both official statistics and information from job aggregators. While such platforms are not employment-focused, the sheer number of active users generates an incentive for users to search for jobs through their online social network, leveraging social capital from their weak ties (Granovetter 1973; Burke and Kraut 2013). In turn, companies are responding by increasingly advertising job vacancies on these platforms (Bhanot 2012). Compared to spe-

cialized platforms such as LinkedIn, job postings on general-purpose social media platforms may provide a greater diversity of sources, from job aggregators to small businesses through individuals advertising jobs. Finally, some of these generalist platforms, and in particular Twitter, have historically provided researchers with data access, enabling such research. Despite the potential informational value of this data source, there is only limited research on the subject. Prior work mostly sought to identify vacancies from social media posts (Tonneau et al. 2022) without categorizing or aggregating the occupational categories and locations these vacancies cover. In this context, the characteristics of job postings on general-purpose platforms remain largely unknown.

In this work, we aim to bridge this gap by evaluating the potential of Twitter as a data source for estimating U.S. labor market demand. We focus on the U.S. market as there are multiple sources of readily available official statistics about the U.S. labor market, enabling us to evaluate our methods comprehensively. To that end, we collect a large dataset of tweets that were posted between 2015 and late 2022 by users with profile locations in the U.S. We then apply an existing classifier to identify tweets containing information on job openings (Tonneau et al. 2022), and train a BERT-based Named Entity Recognition (NER) model to extract job location and occupation information from the tweets. With this information, we create a Twitter-based index of job vacancies by U.S. city and occupational class, and characterize its coverage and representativeness using official statistics from the Job Openings and Labour Turnover Survey (JOLTS) and the Occupational Employment Statistics (OES). We also compare our index with an existing aggregate index of online job postings, finding that our index is better aligned with official statistics for some sectors. Finally, we find that using our index can consistently improve the prediction of the employment rate across occupations and states.

Thus, the current study has the following contributions:

- A methodology for accurately extracting occupation and location information from job postings on social media.
- A quantitative evaluation of the coverage and representativeness of our Twitter-based job vacancy index, including a comparison with a popular index from a large job postings aggregator.
- Empirical evidence showing small but consistent improvements in using our Twitter index for macroeconomic forecasting across occupations and states.

Related Work

Previous research that used social media data to make inferences about the labor market has primarily focused on creating indices of labor market activity through keyword analysis (Antenucci et al. 2014) as well as non-textual information such as diurnal rhythms and mobility patterns (Llorente et al. 2015). Recent work has also leveraged pre-trained language models to identify job postings with a higher accuracy. In particular, Tonneau et al. (2022) trained a BERT-based binary classifier using Active Learning and crowdsourced labels to accurately identify job-related tweets. To the best of our knowledge, the present work is the first to examine the characteristics of job postings on social media and compare them to a broader range of online and offline job postings across occupations and geographic areas.

Another line of related work pertains to information extraction from job posting ads. Such prior work includes skill and requirement extraction (Bhola et al. 2020; Wild et al. 2021; Zhang et al. 2022), knowledge base creation (Van Hautte, Schelstraete, and Wornoo 2020; Bana et al. 2020), job offer classification (Gnehm and Clematide 2020; Zhang et al. 2019), and the use of neural networks to learn text similarities (Decorte et al. 2021; Neculoiu, Versteegh, and Rotaru 2016). An important contrast from social media posts is that job posting ads are often structured and more detailed. Identifying the characteristics of job vacancies from short tweets, for example, is especially challenging due to their brevity, informality, limited context information, and the absence of existing labeled datasets for this task (Nair and Shetty 2015). Due to these differences, it is not clear that models used for extracting information from job posting ads can simply be ported to social media content.

Existing approaches to extract structured information from tweets include regular expressions (Antenucci et al. 2014; Pano and Kashef 2020; Middleton 2015) and Named Entity Recognition (NER) (Liu and Zhou 2013; Liu et al. 2011, 2013). Regular expressions only identify predefined patterns, which may fail to capture the diversity of ways job information could be presented. Similarly, off-the-shelf NER methods may be unable to accurately identify job-related entities such as occupational class due to the informal and noisy nature of tweets. Several studies attempted to overcome these challenges with Convolutional Neural Networks (CNN) (Izbicki, Papalexakis, and Tsotras 2019), language modeling (Ponte and Croft 2017), and Part of Speech (POS) tagging techniques (Derczynski et al. 2013). However, these studies relied on additional information, including social network data or the accumulation of tweets from the same user to improve prediction accuracy. For our task, the information about advertised vacancies may be completely unrelated to the characteristics of the user who posted it. Therefore, we focus on extracting information from the tweets themselves.

To accurately extract information from social media text, our proposed methodology builds on the advancements in large language modeling in recent years, specifically the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al. 2019). BERT is a pre-trained language model that can be fine-tuned for Named Entity Recognition, allowing us to identify job-related entities in unlabeled tweets without relying on auxiliary information. Our approach is the first to utilize BERT for extracting job-related entities from tweets in this manner. BERT has proven to be effective in extracting entities from informal and noisy text (Nguyen, Vu, and Tuan Nguyen 2020), making it well-suited for the nature of tweets. Its architecture has similar elements to emerging GPT-based models, which may improve performance even further.

The literature also highlighted important caveats to con-

sider when using social data or "big data" for time series prediction. In particular, the failure of the Google Flu Trends prediction provides a cautionary tale for making predictions that are highly dependent on features selected solely based on temporal correlations, which assume that user or algorithmic behavior does not change over time, or validate predictions against a single source of information (Lazer et al. 2014). In this context, Jungherr, Jürgens, and Schoen (2012) highlight the importance of articulating a clear analysis plan that is free of arbitrary inclusion/exclusion criteria. It is also critical to consider potential biases between the social media sample and the actual population being studied (Ruths and Pfeffer 2014). In this regard, one of the earliest studies of Twitter's representativeness examined the prevalence of different occupation categories (Mislove et al. 2011). They found that individuals in creative occupations, such as artists, designers, and writers, are over-represented while individuals in more traditional occupations, like manufacturing and transportation, are underrepresented. The present study builds upon this knowledge by taking a principled approach to building our index, starting with careful modeling and extraction of key pieces of information about job vacancies, triangulating the findings against multiple sources of information, directly evaluating representativeness, and assessing predictive ability of the new index jointly with other existing signals.

Data

Vacancy Tweets

The primary dataset used in this work is a set of 22 million tweets that we have identified as describing job vacancies with occupation information, U.S. location, and without duplication. To derive this dataset, we identify users with a profile location in the U.S. in the Twitter Decahose, a 10% random sample of all tweets, between 2010 and 2019. Then, we collect users' timelines until December 2022 using the Twitter API, and use user mentions to expand the set of U.S.-based users for whom we collect timelines.

We then apply the binary BERT-based classifier developed by Tonneau et al. (2022) on all the English tweets in our dataset. We use a score threshold that corresponds to an average precision of 0.9, and exclude retweets to reduce the occurrence of the same job posting multiple times. This results in a set of 173 million tweets that are likely to contain job vacancies. However, this set still contains duplicate postings and vacancies that lack one or more critical pieces of information such as location or occupation.

We adapt the deduplication approach of Zhao, Chen, and Mason (2021) to remove duplicate job postings. First, we remove any URL or special token from the text. Then, we consider as duplicate tweets that contain the same text (excluding URLs and special tokens), posted by the same user, or contain the same URL within a given month. This deduplication removes about 23% of the tweets each year, and results in a dataset with over 133 million tweets (Fig. 2). Then, applying the NER model developed in this work and detailed in the Methods section, we obtain the final set of 22 million tweets with occupation information and location in the U.S. that are not duplicates of other postings. In this final dataset, most tweets have links to external sites (85.61%) but notably not all of them. The average length of a tweet in our dataset is 16.9 words long with a median of 16.5 words.

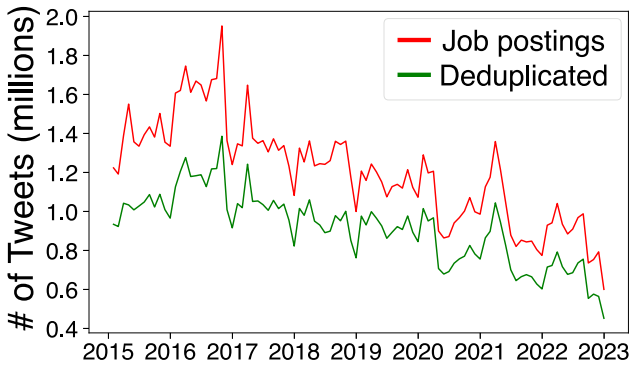

Figure 2: Monthly number of tweets (millions) containing job postings before deduplication (in red) and after deduplication (in teal).

Job Occupations

We collect job occupations from three different datasets in order to maximize linguistic diversity. In all three datasets, job occupations are categorized using the International Standard Classification of Occupations 2008 code (ISCO-08), a widely used hierarchical classification of occupations into increasingly narrow categories (see Fig. 3).

The first dataset is the European Skills, Competences, Qualifications, and Occupations (ESCO) database[1], which contains 3,008 job occupations translated into 28 languages. The occupations are organized in hierarchical form and each occupation is mapped to exactly one ISCO-08 code. The second dataset is the Standard Occupational Classification (SOC)[2], which is commonly used by U.S. federal government agencies for collecting occupational data, enabling comparison of occupations across data sets. It is designed to cover all occupations in which work is performed for pay or profit, reflecting the current occupational structure in the United States, and contains more than 30,000 job occupations. Finally, we use a dataset that JobBERT (Decorte et al. 2021) was trained on for job title normalization. This dataset includes vacancy titles and ESCO codes for over 30,000 unique titles obtained from a large governmental job board.

After removing duplicate entries across all three sources, we obtain a dataset with a total of over 60,000 unique job occupations and their corresponding ISCO-08 classifications in the English language. For the rest of the analysis, we use only the 2-digit (sub-major) ISCO-08 codes (second level of the ISCO-08 codes as shown in Fig. 3) because more granular codes would result in many occupations that were too sparse. We exclude the major category of "Armed Forces Occupations" from the evaluation since it does not appear in the official statistics. We experimented with predicting the

[1] Available at https://esco.ec.europa.eu/en/use-esco.

[2] Available at https://www.bls.gov/soc.

2-digit, 1-digit, and a combination of the two as detailed in the Results section.

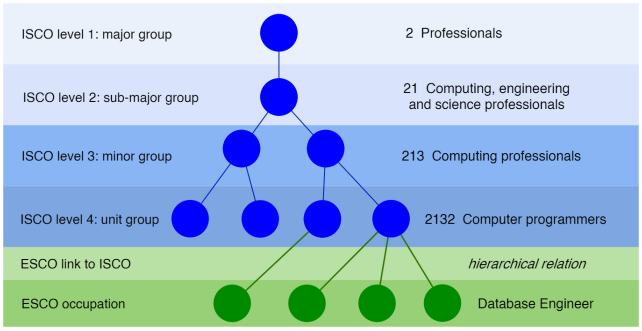

Figure 3: Examples of ISCO-08 classification of occupations listed in the ESCO dataset.

Dataset	Occupation	Period	Freq.	Geo.
BGT (N=90)	Major	2010 - 2019	Year	National
OES (N=4,590)	Major	2010 - 2019	Year	City / State
CPS (N=98,523)	Sub-Major	2015 - 2022	Month	State
JOLTS (N=4,284)	N/A	2015 - 2022	Month	State
Twitter (N=90,226)	Sub-Major	2015 - 2022	Month	City / State

Table 1: Summary of datasets used with occupational granularity matching ISCO-08 levels, coverage years, temporal granularity (yearly or monthly frequency), and geographic granularity. Sample size (N) indicates the number of data points in the time series.

Labor Market Statistics

To evaluate the coverage and representativeness of our new Twitter index, we collect information from two data sources: official statistics and an alternative index of online job postings. Although some of these sources provide estimates that are only related to job openings and none of them has the granularity of our Twitter index (monthly, sub-national, per occupation), it allows us to evaluate different aspects of our index.

We first collect official labor market statistics from the Job Openings and Labor Turnover Survey (JOLTS), the Occupational Employment Statistics (OES) program, and the Current Population Survey (CPS). The JOLTS survey is conducted monthly by the U.S. Department of Labor's Bureau of Labor Statistics (BLS). It provides the number of job openings, hires, and total separations (quits, layoffs, and discharges, and other separations) at the state level. The OES program provides annual employment rates (as well as wage estimates) for 830 occupations at the national, state, metropolitan, and city levels. OES classification is only compatible with ISCO-08 1-digit codes. We also collected official statistics from CPS (a survey conducted by BLS), which has monthly employment rate information aggregated by state and 2-digit ISCO-08 occupational category. Although CPS has the closest granularity to our Twitter index, it describes employment rates, which are related to available jobs in the market but are still different.

Second, we use an alternative index of online job postings from Burning Glass Technologies (BGT), which is one of the largest aggregators of online job postings. Since BGT data is not publicly available, we rely on data released by prior work (Cammeraat and Squicciarini 2021) containing the annual number of job openings per occupation using the 1-digit ISCO-08 codes.

In terms of timeframe, JOLTS and CPS data cover the years 2015 to 2019, while BGT and OES cover 2010 to 2019. Since we are using ISCO-08 classification and the CPS data uses a different classification, we use crosswalks to map from one to another, following the change introduced in this mapping in 2018. A summary of the datasets used in this work can be seen in Table 1.

Methods

Our methodology consists of three parts. First, we train an NER model to extract job occupation and location information from tweets describing job vacancies. Then, we develop two models: one classifier that learns the association between occupation names and ISCO-08 categories, and a separate model that maps locations from text to U.S. cities. Finally, we aggregate the resulting data over time to create a Twitter index of job openings per city and occupational class, and evaluate the ability of this new index to improve the predictability of official statistics.

Information Extraction From Job Postings

We formulate the extraction of job occupation and location information from job posting tweets as a NER task. We randomly sample 3,500 tweets from our dataset of vacancy tweets and annotate them for text spans (using the standard Inside, Outside, Beginning tagging scheme) that represent job occupations (OCC), locations (LOC), organizations (ORG), or none of the above (MISC). Figure 1 provides an example of a tweet with the location and occupation entities marked, highlighting the required entities for inclusion in our index. A single annotator labeled the data, with a second annotator labeling a sub-sample of tweets to ensure the labeling quality. A total of 12,235 entities were tagged, which were then used to fine-tune a Conversational BERT model for the NER task (Burtsev et al. 2018). We used up to 10 epochs for training, maximizing accuracy. We evaluated the performance using standard 10-fold cross-validation, and compared the model against several baselines: Multinomial Naive Bayes (MNB), vanilla BERT (BERT-base-cased), and JobBERT (Decorte et al. 2021). For MNB, we used a Bag-of-Words representation to turn tweets into vectors of word counts for the model. The MNB model was chosen due to its simplicity and flexibility, as well as its good performance on various NER benchmarks (Khan et al. 2016). Vanilla BERT

enables the evaluation of a larger and more recent language model, while JobBERT enables the assessment of the contribution of domain knowledge as it was further pre-trained on job postings.

Mapping Job Occupations to ISCO Codes

Next, to map the job occupations to ISCO-08 codes, we train a model using the job occupations dataset presented in the Data section. Since the text describing the job occupations is rather short, we utilize the Sentence BERT architecture (Reimers and Gurevych 2019) and specifically the "all-mpnet-base-v2" model to generate sentence embeddings for job occupations. These embeddings are then used as features to train a model to predict the ISCO-08 category of a job occupation. We experiment with various deep learning models and evaluate their performance using the Mean Reciprocal Rank (MRR), which emphasizes assigning a high rank to the correct class. Since the correct label is known, after sorting the predictions according to their score, we can identify the rank of the correct label.

To test the granularity of our job occupational classification, we train and evaluate our models on three variations of the ISCO-08 categorization: 1-digit, 2-digit, and a 2-digit collapsed variant. The first two variations use the major and sub-major categories of ISCO-08, while the third version merges some sparse classes under a major category to effectively increase the training size. We perform hyper-parameter tuning on the learning rate, dropout rate, batch size, and regularization to maximize the MRR. Our best model had a learning rate of $1e{-}4$, batch size of 512, dropout rate of 0.2, and no regularization. We evaluate this model against the baselines using 10-fold cross-validation, and use this best-performing model to infer the ISCO-08 category of job occupations detected by the NER model in our dataset of vacancy tweets.

Location Classification

To obtain a comprehensive list of locations in the United States, we use the World Cities Database[3] which is a large listing of cities and includes 140,000 US cities. We employ the Sentence BERT model called "all-mpnet-base-v2" to compute embeddings for the combined city and state names, which we then map to latitude and longitude coordinates. To map locations extracted from tweets to geo-locations, we first use the NER model developed in the preceding section to identify all spans of tokens predicted as locations (LOC). We then convert these strings to sentence embeddings and use cosine similarity to map the extracted location to a known location in the World Cities Database. As most locations are abbreviated versions of city and state names, we also compare the embedding-based approach to a simple fuzzy matching using the Levenshtein distance.

Time Series Prediction

In order to evaluate the predictive ability of our Twitter index, we use it to predict official statistics about employment rates from CPS. Our goal is to evaluate whether our Twitter index can consistently improve the predictive accuracy of a simple auto-regressive model. More formally, our time series prediction model is the following:

$$X_s(t) = \alpha_s + \sum_{i=1}^{p} \beta_{s,i} X_s(t-i) + \sum_{i=1}^{p} \gamma_{s,i} T_s(t-i) + \epsilon_s \tag{1}$$

where $X_s(t)$ is the CPS time series for each state s, p is the order of the auto-regressive model, $T_s(t)$ is our Twitter index for each state, $O_{s,j}$ represents a dummy variable for the occupation j in state s, and ϵ is the error term. α is the intercept term for the model, while β and γ are model coefficients fitted during training. First, we determined the best-performing auto-regressive model by testing different history lengths ($p \in [1, 2, ..., 10]$). Then, using the best-performing history length, we compared the predictive accuracy of the AR model with and without our Twitter index ($\forall i, \gamma_{s,i} = 0$). Our Twitter index, $T_s(t)$, is a normalized number of job-opening tweets by the total number of tweets posted in the state in a given time period. This normalization is important because it allows us to control for changes in the adoption of Twitter, similarly to the employment rates in CPS that normalize for the size of the labor market. For each state s, the first half of each occupation time series is used for training, and the second half is used for testing.

Results

In this section, we report the performance of the different components of our pipeline in identifying and extracting information from job vacancies as well as the overall accuracy of an aggregate Twitter index across occupations, U.S. states, and compared to different indices.

Evaluating Pipeline Components

Information Extraction To evaluate the performance of Conversational BERT in identifying job occupations, locations, and organizations in job posting tweets, we use a standard 70/30 train/test split of the annotated IOB dataset we created. We compare our model with a MNB, a vanilla BERT Base and JobBERT, using standard performance metrics (Accuracy, F1, Precision, Recall) in classifying the different tags (job occupation, location, organization, and miscellaneous). These results are presented in Table 2. We find that Conversational BERT achieves high accuracy for all tags, especially for occupations and locations, which are plausibly easier classification tasks due to the use of capitalization, and repeated linguistic constructs. In contrast, identifying organizations (ORG) is considerably more challenging, possibly due to the diversity of company names and lack of persistent structure.

While BERT-based models consistently outperform the MNB model, Table 2 shows that the performance is comparable across BERT models. The gains of BERT-based models are consistent across all performance metrics and tags, except for MISC, with gains over MNB being particularly large for organizations. The lower performance of MNB in

[3]https://www.kaggle.com/datasets/max-mind/world-cities-database

	Base model	OCC	LOC	ORG	MISC
Accuracy	CBERT	**0.97**	**0.98**	**0.9**	**0.88**
	JobBERT	0.93	0.94	0.89	0.86
	BERT-BASE	0.95	0.96	0.87	0.87
	MNB	0.77	0.79	0.35	0.85
F1	CBERT	0.79	0.88	0.63	0.63
	JobBERT	0.79	0.88	0.68	0.58
	BERT-BASE	**0.81**	**0.89**	**0.7**	0.72
	MNB	0.64	0.82	0.44	**0.87**
Precision	CBERT	**0.85**	0.89	**0.78**	0.62
	JobBERT	**0.85**	**0.93**	0.69	0.57
	BERT-BASE	0.8	0.89	0.64	0.61
	MNB	0.54	0.86	0.57	**0.89**
Recall	CBERT	**0.83**	**0.93**	0.7	0.72
	JobBERT	0.74	0.83	0.69	0.58
	BERT-BASE	0.82	0.9	**0.76**	0.72
	MNB	0.77	0.79	0.35	**0.85**

Table 2: The performance of extracting job occupation (OCC), location (LOC), organization (ORG), and miscellaneous (MISC) tokens from job posting tweets by three different BERT-based models and a Multinomial Naive Base model (MNB). CBERT being Conversational BERT.

identifying the ORG tag may be attributed to the requirement of drawing contextual information for accurate classification. The MNB model outperforms the other models for MISC, which often includes diverse and miscellaneous entities that may not exhibit consistent linguistic patterns. BERT models, which heavily rely on learning contextual representations from large-scale text data, may have difficulty capturing the specific patterns and nuances associated with the "MISC" tag. The different BERT models — without any additional pre-training (BERT-Base), with additional pre-training on job postings (JobBERT), and with additional pre-training on social media data (Conversational BERT) — all perform within a few percentage points of one another. This suggests that for the task of extracting information from job posting tweets, additional training on social media data or on job posting ads does not bring significant gains. The gains are more likely to come from the fine-tuning on the NER task.

Job Occupational Classification Having successfully extracted job occupations from job posting tweets, we train a classifier to map each occupation to a ISCO-08 category. We experiment with different levels of granularity and model hyper-parameters optimized using a validation set with a 60%/20%/20% train/validation/test splits.

Table 3 reports the performance of our best-performing model, obtained after 38 training epochs, in terms of MRR, Recall at 1 (R@1), Recall at 5 (R@5), and Accuracy (Acc) for different granularity levels of the ISCO-08 categorization. At the Major ISCO-08 level containing 10 categories, we find that our model exhibits very good performance, with an average MRR of 0.83 corresponding to having the correct label placed at an average rank of 1.2. When considering the Sub-Major ISCO-08 level with 43 categories, the MRR decreases to 0.75 but the correct category is still ranked between first and second place on average. The Sub-Major Variant (SMV), for which we collapse sparse categories into larger ones, improves performance relative to the Sub-Major level but still performs worse than with Major levels. Recall values (and accuracy) show that the correct category is ranked first about 70% of the time, within the top five results roughly 90% of the time, and the performance is balanced across categories as indicated by high average macro values.

ISCO-08 level		Major	SMV	Sub-Major
MRR	Average	0.83	0.79	0.75
R@1	Macro	0.72	0.69	0.61
	Micro	0.72	0.68	0.63
R@5	Macro	0.97	0.92	0.81
	Micro	0.98	0.94	0.91
Acc	Macro	0.72	0.89	0.86
	Micro	0.72	0.68	0.63

Table 3: The performance of classifying extracted job occupations from tweets into ISCO-08 Major (10 categories), Sub-Major (43 categories), or Sub-Major variant with sparse categories collapsed (SMV; 23 categories).

Location Tagging To evaluate our mapping from extracted locations to geo-located cities, we label the correct location of 1,900 random job posting tweets containing a predicted location. The location classification is at the city level and for this reason, our annotation process is limited to city-level identification. Tweets that did not contain a city or had multiple locations are discarded. We compare the matching performance by semantic similarity using sentence embeddings and fuzzy matching based on Levenshtein distance. Cosine similarity on sentence-BERT embeddings achieves an accuracy of 98.4%, overperforming fuzzy matching based on Levenshtein distance which achieved an accuracy of 93.4%.

Representativeness of Twitter Job Opening Index

Geographical Representativeness We assess the geographical representativeness of our Twitter index by comparing the number of job openings on Twitter with the number of job openings from the Job Openings and Labor Turnover Survey (JOLTS) in each U.S. state. We normalize the number of job openings in each state by expressing them as the share of the total job openings across all states. The normalized value for state i, denoted N_i, is calculated using:

$$N_i = \frac{T_i}{\sum_{j=1}^{n} T_j} \tag{2}$$

where T_i is the number of job openings in state i, and n is the total number of states.

Figure 4 shows the normalized number of job postings on Twitter (X-axis) and job openings on JOLTS (Y-axis) where each point is a state. Values were normalized using Eq. 2. The three panels in the figure provide increasingly zoomed-in versions, with the left panel showing all states, the middle panel focusing just on the green-marked area in the left panel, and the right panel showing the further zoomed-in area marked in red in the middle panel. In each panel, a

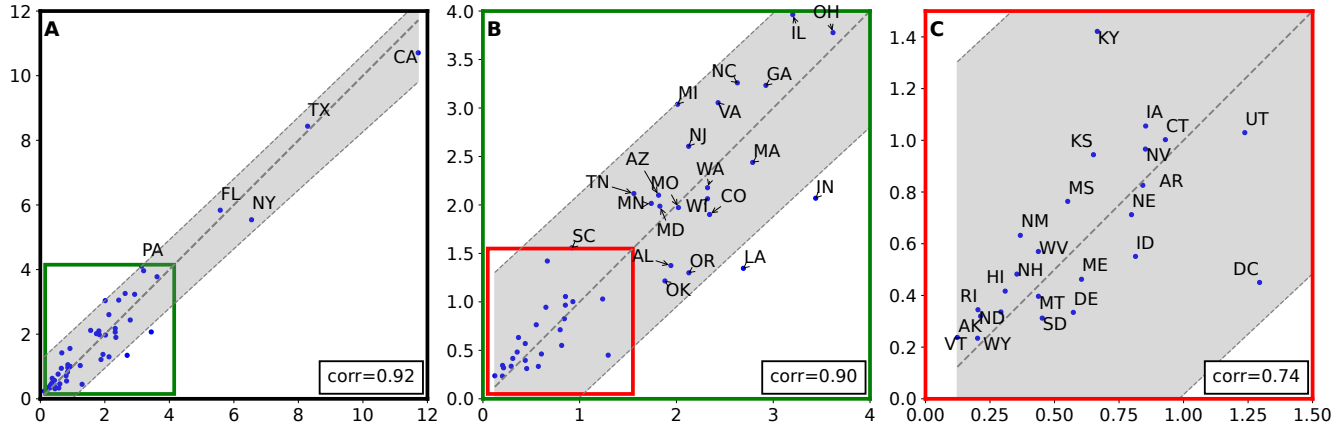

Figure 4: Geographical Representativeness of Twitter at the state level. The x- and y-axes show the percentage of job postings on Twitter and the percentage of job openings on JOLTS per state, respectively. The three panels provide increasingly zoomed-in versions, where the left panel (A) shows the data for all states, the middle panel (B) corresponds to the green-marked area in the left panel, and the right panel (C) shows the area marked in red in the middle panel. The dashed diagonal line ($y = x$) designates perfect alignment between the two indices with the gray shaded area around it showing the confidence interval (using one standard deviation) of a linear regression model fitted to data. The Spearman Rank Correlation coefficient is shown in the bottom right side in each panel. ($p < 0.001$ in all panels)

dashed diagonal line shows the line of perfect alignment between the Twitter and JOLTS indices ($y = x$), and a shaded gray area shows one standard deviation of a linear regression model fitted to the data. The Spearman Rank Correlation coefficient is noted below each panel based on the set of data points included in the panel, i.e., the left panel reports the correlation across all states.

We observe a strong and significant positive correlation between our Twitter and JOLTS at the state level (Figure 4). One can see that many points are close to the identity line and within one standard deviation off of the mean. This is further supported by the high rank correlation coefficients, ranging from 0.92 when considering all states to 0.74 when considering the smallest subset of states shown in the right panel (all with $p < 0.001$). Focusing on a few of the points that deviate the most from the identity line, we observe that California and New York are over-represented on Twitter, which may be explained by the high penetration rate of Twitter in these states. Additionally, Indiana and Louisiana are also over-represented on Twitter, but further examination reveals that this deviation is not consistent over time and is mostly driven by a high number of job postings posted on Twitter in the years 2015 and 2016. Pennsylvania and Kentucky stand out as relatively underrepresented in our Twitter index, but still within the margin of error. Next, we examine the representativeness of our Twitter index across occupations and time.

Occupational Representativeness We assess the occupational representativeness of our Twitter index by comparing the per occupation number of job openings from Twitter and the number of job openings from BGT, relative to the official employment rate from OES. We obtain the share of job openings on Twitter and separately the share of job openings in BGT per occupation and year by normalizing the raw count numbers using the same formula as for the Geographical Representativeness, with the only difference being that T_i is now the number of job openings in occupation i, and n is the total number of occupations. To compare these normalized indices to official employment rates, we calculate the difference between the Twitter and BGT indices on the one hand and the employment rate obtained in the same occupation and year on the other hand. Therefore, positive values represent an over-representation of an occupation in the index relative to the official employment statistics, and negative values designate an under-representation.

Figure 5 shows those differences in Occupational Representativeness for our Twitter index and BGT across nine major ISCO-08 categories over time. The y-axis shows the difference of these indices from the employment rate. Each panel represents a 1-digit (major) ISCO-08 category with a dashed black line at $y = 0$ representing the employment rate in that occupation category. Lines closer to zero indicate closer alignment, and therefore better representation, of the index relative to employment rates. Above- and below-zero values correspond to over- and under-representation of the index, respectively, compared to official employment rates. It should be noted that the scale on the y-axis varies considerably across subplots as some occupations only slightly differ from the official statistics and some occupations differ from the official statistics by double-digit percentage points.

We observe that the representativeness of the Twitter and BGT indices vary across occupation categories (Figure 5). Twitter is more representative in occupational categories for "Technicians and associate professionals" and "Craft and related trades workers", while BGT over- and under-estimates these categories, respectively. Conversely, BGT aligns more closely with official employment rates

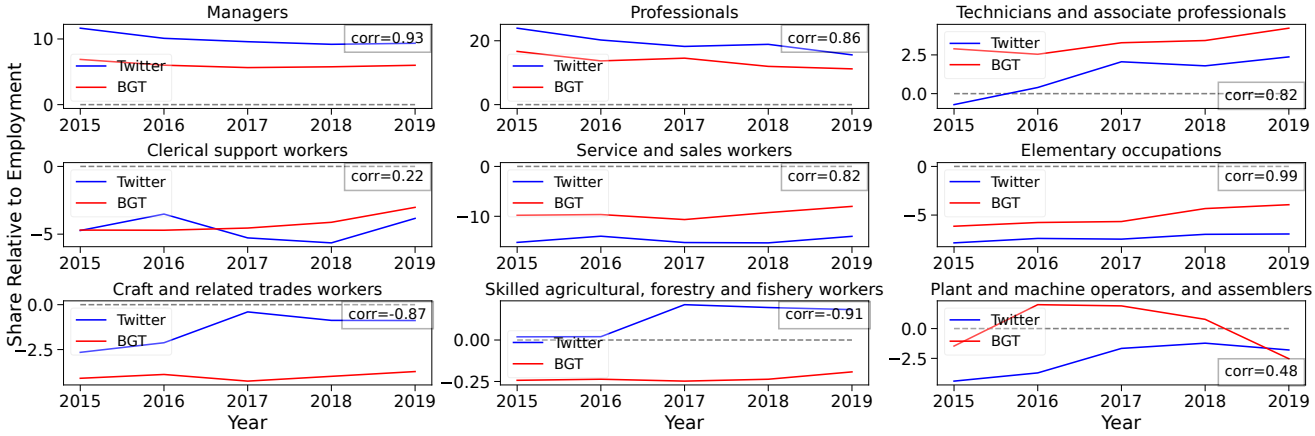

Figure 5: Occupational representativeness of Twitter (blue) and BGT (red) over time. The X-axis shows time (years), and the Y-axis shows the share of job openings as a difference from official employment rates per occupation (obtained from OES). Occupations are aggregated using 1-digit ISCO-08 codes. The share of job openings per occupation is calculated separately for Twitter and BGT before deducting the annual employment rate. Therefore, above-zero or below-zero values indicate over- or under-representation, respectively. Zero values (highlighted by dashed black lines) represent the employment rate per occupation over time. Pearson correlation is shown in each panel.

in several occupation categories ("Managers", "Professionals", "Service and sales workers", and "Elementary occupations"). In these categories, our Twitter index has an over-representation of "Managers" and "Professionals", and an under-representation of "Service and sales workers" and "Elementary occupations". For five out of the nine occupations, there is a strong correlation of 0.8 or more between the Twitter index and BGT. Notably, this strong correlation holds not only in categories that are generally associated with high-skill work (e.g. "Professionals"), but also in categories associated with low-skill work (e.g. "Elementary occupations").

Employment Prediction Figure 6 shows the percentage improvement in Root Mean Square Error (RMSE) obtained by incorporating the Twitter index into a basic AR(1) model for each state and occupation as specified in Equation 1. We use an AR(1) model because it produced the best Akaike information criterion (AIC) across all model orders (p values) that we tested. Occupations are aggregated at the 2-digit ISCO-08 code level. The model was fitted using monthly data from the first half of the study period (2015-2018) and was tested on the second half of the period (2019-2022). The warmer the color of the grid cell in the figure, the closer the model predictions were to the official statistics compared to a model without our Twitter index. States and occupations that are absent in the Twitter index are shown in black in the corresponding cells.

Our results demonstrate that including the Twitter index improves the predictions for most states and occupations, with an average improvement of 0.85% across occupations and 0.73% across states. Occupations with the best improvements are "Assemblers", "Stationary Plant and Machine Operators", "Market-oriented Skilled Agricultural Workers", and "Handicraft and Printing Workers" in the states where they are present. However, we observe a few exceptions, such as Alabama, where fluctuations in the employment rate cause a decrease in RMSE, making the presence of the Twitter index more significant than other states such as Texas. Furthermore, the index achieved substantial improvements in RMSE for the "Skilled agricultural, forestry and fishery workers", "Craft and related trades workers", and "Plant and machine operators, and assemblers" categories. This is in line with results from the previous section showing that Twitter represents these categories better than BGT.

Discussion

In this study, we developed a methodology for extracting job occupations and location information from tweets containing job openings. By fine-tuning and evaluating the components of our extraction pipeline (NER, mapping job occupations to ISCO categories, geographic mapping), we derived an aggregate index of available jobs solely based on social media data. We found that the newly-constructed index is strongly correlated with JOLTS, an official statistic of job openings from the U.S. Bureau of Labor Statistics. We also found that in some sectors, our index is better aligned with official statistics than an existing aggregate index of online job postings. Moreover, we showed that using our index in a simple auto-regressive model can consistently improve the prediction accuracy of employment statistics across states and occupations, indicating that the signal extracted from social media can complement more traditional sources of information about the labor market.

Our work provides promising results along multiple dimensions. First, as a statistical approach, our methodology provides a complementary but consistent estimate of official statistics that can improve estimation and prediction. This can improve the estimation in well-resourced countries

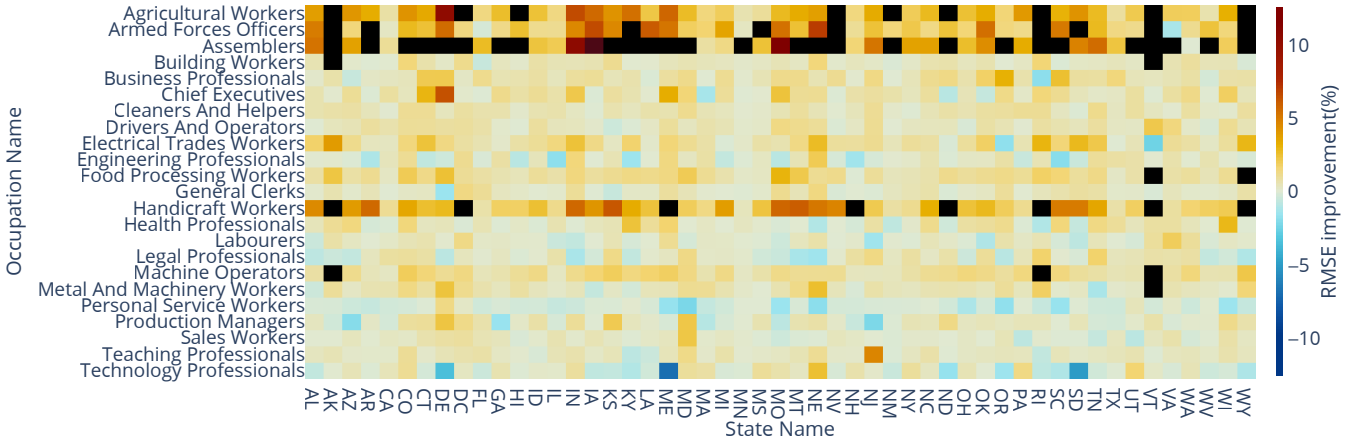

Figure 6: Percentage improvement in RMSE when including our Twitter Index in an AR(1) model for different occupations (rows, 2-digit ISCO level) and U.S. states abbreviations (columns). Warmer colors (red or yellow) indicate improvement over the baseline, while cooler colors (blue or teal) show worse predictive performance. Black tiles represent combinations of state and occupation that are missing in our Twitter index.

with extensive surveying infrastructure — as we showcased for the U.S. — and can potentially aid surveyors in allocating resources more efficiently in areas and times where changes are likely to happen. An even greater potential lies in the application of this approach in low-resource countries where economic surveying is either absent, severely lacking, or lagging. In such countries, especially where social media out-paced surveying infrastructure, the proposed social media-based index can provide the most recent measure of labor market demand. Moving beyond estimation, the ability to detect and classify job vacancies through social media can assist in matching job seekers with relevant employment opportunities. This can also help organizations offering training programs to adapt and target their programs better. In addition, as demonstrated by the comparison with BGT, this approach can capture low-skill and high-skill jobs, which is important for a more comprehensive and *inclusive* measure of the labor market. It remains an open question whether this improved representation relative to job aggregators like BGT stems from the wider adoption of social media by diverse populations, a property of the U.S. market, or an ephemeral characteristic of existing job posting sites. Regardless, these different sources of information can be combined into a more comprehensive index. Finally, our approach resulted in a small but consistent improvement in prediction accuracy across sectors and states using relatively simple aggregation and time-series prediction methods. Surely, further improvements can be attained using more sophisticated stratification methods, better deduplication methods (e.g., using information from outside links), and more complex time-series models that include other factors known to affect the availability of jobs (e.g., inflation rate).

Another important avenue for future research is to expand our approach to other large-scale social media platforms (e.g., Facebook, Reddit) and to other languages. While our approach is statistical and can be generalized to other platforms, differences in affordances and user base make it difficult to predict how well this approach will work on other social media platforms. It is possible that most job postings on Facebook take place in non-public groups, and therefore, public and private content would need to be considered separately. Similarly, the anonymous nature of Reddit can potentially impact the type of jobs users post on the platform. Therefore, it is essential to carefully consider both who is using the platform and how they are using it to post about jobs in each context to get the most accurate estimates. Moreover, to fully materialize the potential of our approach in low-resource countries, one must extend beyond the English language. Such an approach can potentially substitute our BERT-based models with multilingual variants, most likely with some performance degradation. Finally, another important aspect that future work could explore is accounts that are redistributing the job offers. It is interesting to understand who is sharing these offers, whether they are automated accounts, what content is being shared, and who are the main sources of these jobs.

Our study has several methodological limitations that can be improved in future work. We did not identify posts advertising multiple vacancies, which requires closer attention to the particular linguistic forms used to describe them. We did not directly model the duration of availability of job postings, which could affect the accuracy of our index. A better model could potentially use information from job posting sites to deduct jobs that are no longer available and adjust the model per sector. We considered adversarial manipulations of the index as outside the scope, but future work could aim to develop estimates that are resilient to such attacks. Lastly, we note that there are areas where the official statistics are inherently problematic as ground truth and corrections are non-trivial. For example, our index may capture job vacancies that are not formally advertised, while official statistics may miss these vacancies. The gaps may be particularly large in the informal sector or when work is undocumented.

Finally, the success of our approach and its ability to deliver public social good is dependent on the availability of large-scale social media data. Unfortunately, recent changes in Twitter's API pricing and availability for academic research have severely hampered the ability to use Twitter data for macroeconomic decision-making. As academic researchers, we are concerned about the diminished capacity of our field, and by extension of the public, to observe and draw insights from social media platforms. Nevertheless, we believe that our methodology can be applied to other platforms with some care. We hope our work motivates further collaborations between platforms and government as well as national statistical institutes to provide access to the necessary data.

Code Availability Statement

All replication code and fine-tuned models are publically available at https://github.com/Socially-Embedded-Lab/twitter-job-postings for academic usage. Aggregate state and occupation data is available for replication purposes. No tweet- or user-level information is provided to protect individual users' privacy.

Ethical Statement

This study analyzed publicly available data in an aggregate manner. Except for the initial filtering of profiles based on U.S. locations, no inference was conducted at the user level, and no member of the research team examined individual accounts. We strongly advise additional calibration and testing before using our methodology for economic decision-making in other contexts not directly tested in this study.

Following scientific best practices, we disclose that we have no competing interests that may impact the results or interpretation of our research. We affirm that our objective is to present reliable and informative findings that contribute to a better understanding of labor market dynamics. No financial or personal relationships exist that could influence our research or its conclusions.

References

Antenucci, D.; Cafarella, M.; Levenstein, M.; Ré, C.; and Shapiro, M. D. 2014. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research.

Bana, S.; Brynjolfsson, E.; Rock, D.; and Steffen, S. 2020. job2vec: Learning a representation of jobs. *Stanford Digital Economy Lab*.

Bhanot, S. 2012. Use of social media by companies to reach their customers. *SIES Journal of Management*, 8(1).

Bhola, A.; Halder, K.; Prasad, A.; and Kan, M.-Y. 2020. Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5832–5842. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Burke, M.; and Kraut, R. 2013. Using Facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1419–1430.

Burtsev, M.; Seliverstov, A.; Airapetyan, R.; Arkhipov, M.; Baymurzina, D.; Bushkov, N.; Gureenkova, O.; Khakhulin, T.; Kuratov, Y.; Kuznetsov, D.; Litinsky, A.; Logacheva, V.; Lymar, A.; Malykh, V.; Petrov, M.; Polulyakh, V.; Pugachev, L.; Sorokin, A.; Vikhreva, M.; and Zaynutdinov, M. 2018. DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of ACL 2018, System Demonstrations*, 122–127. Melbourne, Australia: Association for Computational Linguistics.

Cammeraat, E.; and Squicciarini, M. 2021. Burning Glass Technologies' data use in policy-relevant analysis: An occupation-level assessment. *OECD Science, Technology and Industry Working Papers*.

Carnevale, A. P.; Jayasundera, T.; and Repnikov, D. 2014. Understanding online job ads data. *Georgetown University, Center on Education and the Workforce, Technical Report (April)*.

Decorte, J.-J.; Van Hautte, J.; Demeester, T.; and Develder, C. 2021. JobBERT : understanding job titles through skills. In *FEAST, ECML-PKDD 2021 Workshop, Proceedings*, 9.

Deming, D.; and Kahn, L. B. 2018. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1): S337–S369.

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the international conference recent advances in natural language processing ranlp 2013*, 198–206.

Devarajan, S. 2013. Africa's statistical tragedy. *Review of Income and Wealth*, 59: S9–S15.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Gnehm, A.-S.; and Clematide, S. 2020. Text Zoning and Classification for Job Advertisements in German, French and English. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 83–93. Online: Association for Computational Linguistics.

Granovetter, M. S. 1973. The strength of weak ties. *American journal of sociology*, 78(6): 1360–1380.

Izbicki, M.; Papalexakis, V.; and Tsotras, V. 2019. Geolocating tweets in any language at any location. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 89–98.

Jungherr, A.; Jürgens, P.; and Schoen, H. 2012. Why the Pirate Party Won the German Election of 2009 or The Trouble

With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment". *Social Science Computer Review*, 30(2): 229–234.

Khan, W.; Daud, A.; Nasir, J. A.; and Amjad, T. 2016. A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait journal of Science*, 43(4).

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The parable of Google Flu: traps in big data analysis. *science*, 343(6176): 1203–1205.

Liu, X.; Wei, F.; Zhang, S.; and Zhou, M. 2013. Named entity recognition for tweets. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1): 1–15.

Liu, X.; Zhang, S.; Wei, F.; and Zhou, M. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 359–367.

Liu, X.; and Zhou, M. 2013. Two-stage NER for tweets with clustering. *Information Processing & Management*, 49(1): 264–273.

Llorente, A.; Garcia-Herranz, M.; Cebrian, M.; and Moro, E. 2015. Social media fingerprints of unemployment. *PloS one*, 10(5): e0128692.

Middleton, S. 2015. Extracting Attributed Verification and Debunking Reports from Social Media: MediaEval-2015 Trust and Credibility Analysis of Image and Video. *Proceedings of the MediaEval 2015 Workshop*.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 554–557.

Nair, L. R.; and Shetty, D. S. D. 2015. Streaming Twitter Data Analysis Using Spark for Effective Job Search. *Journal of Theoretical & Applied Information Technology*, 80(2).

Neculoiu, P.; Versteegh, M.; and Rotaru, M. 2016. Learning Text Similarity with Siamese Recurrent Networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 148–157. Berlin, Germany: Association for Computational Linguistics.

Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14. Online: Association for Computational Linguistics.

Pano, T.; and Kashef, R. 2020. A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. *Big Data and Cognitive Computing*, 4(4): 33.

Ponte, J. M.; and Croft, W. B. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, volume 51, 202–208. ACM New York, NY, USA.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ruths, D.; and Pfeffer, J. 2014. Social media for large studies of behavior. *Science*, 346(6213): 1063–1064.

Tonneau, M.; Adjodah, D.; Palotti, J.; Grinberg, N.; and Fraiberger, S. 2022. Multilingual Detection of Personal Employment Status on Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6564–6587. Dublin, Ireland: Association for Computational Linguistics.

Van Hautte, J.; Schelstraete, V.; and Wornoo, M. 2020. Leveraging the Inherent Hierarchy of Vacancy Titles for Automated Job Ontology Expansion. In *Proceedings of the 6th International Workshop on Computational Terminology*, 37–42. Marseille, France: European Language Resources Association. ISBN 979-10-95546-57-3.

Wild, S.; Parlar, S.; Hanne, T.; and Dornberger, R. 2021. Naïve Bayes and Named Entity Recognition for Requirements Mining in Job Postings. In *2021 3rd International Conference on Natural Language Processing (ICNLP)*, 155–161. IEEE.

Zhang, D.; Liu, J.; Zhu, H.; Liu, Y.; Wang, L.; Wang, P.; and Xiong, H. 2019. Job2Vec: Job title benchmarking with collective multi-view representation learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2763–2771.

Zhang, M.; Jensen, K. N.; Sonniks, S.; and Plank, B. 2022. SkillSpan: Hard and Soft Skill Extraction from English Job Postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4962–4984. Seattle, United States: Association for Computational Linguistics.

Zhao, Y.; Chen, H.; and Mason, C. M. 2021. A Framework for Duplicate Detection from Online Job Postings. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 249–256.

Zhu, T. J.; Fritzler, A.; and Orlowski, J. A. K. 2018. World Bank Group-LinkedIn Data Insights: Jobs, Skills and Migration Trends Methodology and Validation Results. Technical report, The World Bank.

Ethics Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our analysis was conducted on de-identified, aggregate data and no member of the research team examined individual profiles. Our work focused on the U.S. due to the availability of granular official statistics, but the overall goal of this approach is to contribute to accurate estimation of labor market demands with fewer resources.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, this is thoroughly explained in the methods section and is further supported by findings in the results section.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, the issue and use of a non-probability social media sample of labor-related information is thoroughly discussed in the introduction, and the contribution is presented as complementary to existing probability-based measures. This is further expanded in the discussion of limitations at the conclusion of the paper.

(e) Did you describe the limitations of your work? Yes, the limitations are described in the Discussion section.

(f) Did you discuss any potential negative societal impacts of your work? Yes, we see the main negative consequence of this work as misuse of our methods to make inaccurate estimations. To mitigate this potential future misuse, we emphasize in the Ethical Statement section that additional calibration and testing should be conducted before using this methodology for economic decision-making in other contexts not directly tested during this study.

(g) Did you discuss any potential misuse of your work? Yes, as described in the response to the previous question.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we discussed the generalizability of our findings to other social media platforms and languages, and included comments to limit potential misuse.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see "Code Availability Statement" section.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. See Data and Methods sections.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. Statistical results are reported appropriately.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? The models were trained on a single, general-purpose GPU without any specialized requirements and no substantial compute costs were involved in conducting this work.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Methods section.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? We did not discuss misclassification explicitly since the results of our classifications are fed to a time-series model, which more directly evaluates the consequences of misclassification. The Twitter index, despite potential misclassification, consistently and robustly contributes to accurate macroeconomic forecasting.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating / releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? No, but we complied with usage requirements.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, see Code Availability Statement.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because this research only analyzed de-identified, aggregated public data.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, the discussion guided our choice to conduct on de-identified, aggregate analysis of the data, and share only aggregate state and occupation data for replication purposes, not individual tweets. This is stated in the Code Availability Statement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA