Twitter Job/Employment Corpus: A Dataset of Job-Related Discourse Built with Humans in the Loop

Abstract

We present the Twitter Job/Employment Corpus, a collection of tweets annotated by a humans-in-the-loop supervised learning framework that integrates crowdsourcing contributions and expertise on the local community and employment environment. Previous computational studies of job-related phenomena have used corpora collected from workplace social media that are hosted internally by the employers, and so lacks independence from latent job-related coercion and the broader context that an open domain, general-purpose medium such as Twitter provides. Our new corpus promises to be a benchmark for the extraction of job-related topics and advanced analysis and modeling, and can potentially benefit a wide range of research communities in the future.

Introduction

Working American adults spend more than one third of their daily time on job-related activities BIBREF0 —more than on anything else. Any attempt to understand a working individual's experiences, state of mind, or motivations must take into account their life at work. In the extreme, job dissatisfaction poses serious health risks and even leads to suicide BIBREF1 , BIBREF2 .

Conversely, behavioral and mental problems greatly affect employee's productivity and loyalty. 70% of US workers are disengaged at work BIBREF3 . Each year lost productivity costs between 450 and 550 billion dollars. Disengaged workers are 87% more likely to leave their jobs than their more satisfied counterparts are BIBREF3 . The deaths by suicide among working age people (25-64 years old) costs more than $44 billion annually BIBREF4 . By contrast, behaviors such as helpfulness, kindness and optimism predict greater job satisfaction and positive or pleasurable engagement at work BIBREF5 .

A number of computational social scientists have studied organizational behavior, professional attitudes, working mood and affect BIBREF6 , BIBREF7 , BIBREF8 , BIBREF9 , but in each case: the data they investigated were collected from internal interactive platforms hosted by the workers' employers.

These studies are valuable in their own right, but one evident limitation is that each dataset is limited to depicting a particular company and excludes the populations who have no access to such restricted networks (e.g., people who are not employees of that company). Moreover, the workers may be unwilling to express, e.g., negative feelings about work (“I don't wanna go to work today”), unprofessional behavior (“Got drunk as hell last night and still made it to work”), or a desire to work elsewhere (“I want to go work at Disney World so bad”) on platforms controlled by their employers.

A major barrier to studying job-related discourse on general-purpose, public social media—one that the previous studies did not face—is the problem of determining which posts are job-related in the first place. There is no authoritative training data available to model this problem. Since the datasets used in previous work were collected in the workplace during worktime, the content is implicitly job-related. By contrast, the subject matter of public social media is much more diverse. People with various life experiences may have different criteria for what constitutes a “job” and describe their jobs differently.

For instance, a tweet like “@SOMEONE @SOMEONE shit manager shit players shit everything” contains the job-related signal word “manager,” yet the presence of “players” ultimately suggests this tweet is talking about a sport team. Another example “@SOMEONE anytime for you boss lol” might seem job-related, but “boss” here could also simply refer to “friend” in an informal and acquainted register.

Extracting job-related information from Twitter can be valuable to a range of stakeholders. For example, public health specialists, psychologists and psychiatrists could use such first-hand reportage of work experiences to monitor job-related stress at a community level and provide professional support if

necessary. Employers might analyze these data and use it to improve how they manage their businesses. It could help employees to maintain better online reputations for potential job recruiters as well. It is also meaningful to compare job-related tweets against non-job-related discourse to observe and understand the linguistic and behavioral similarities and differences between on- and off-hours.

Our main contributions are:

Background and Related Work

Social media accounts for about 20% of the time spent online BIBREF10 . Online communication can embolden people to reveal their cognitive state in a natural, un-self-conscious manner BIBREF11 . Mobile phone platforms help social media to capture personal behaviors whenever and wherever possible BIBREF12 , BIBREF13 . These signals are often temporal, and can reveal how phenomena change over time. Thus, aspects about individuals or groups, such as preferences and perspectives, affective states and experiences, communicative patterns, and socialization behaviors can, to some degree, be analyzed and computationally modeled continuously and unobtrusively BIBREF12 .

Twitter has drawn much attention from researchers in various disciplines in large part because of the volume and granularity of publicly available social data associated with massive information. This micro-blogging website, which was launched in 2006, has attracted more than 500 million registered users by 2012, with 340 million tweets posted every day. Twitter supports directional connections (followers and followees) in its social network, and allows for geographic information about where a tweet was posted if a user enables location services. The large volume and desirable features provided by Twitter makes it a well-suited source of data for our task.

We focus on a broad discourse and narrative theme that touches most adults worldwide. Measures of

volume, content, affect of job-related discourse on social media may help understand the behavioral patterns of working people, predict labor market changes, monitor and control satisfaction/dissatisfaction with respect to their workplaces or colleagues, and help people strive for positive change BIBREF9 . The language differences exposed in social media have been observed and analyzed in relation to location BIBREF14 , gender, age, regional origin, and political orientation BIBREF15 . However, it is probably due to the natural challenges of Twitter messages — conversational style of interactions, lack of traditional spelling rules, and 140-character limit of each message—we barely see similar public Twitter datasets investigating open-domain problems like job/employment in computational linguistic or social science field. Li et al. li2014major proposed a pipelined system to extract a wide variety of major life events, including job, from Twitter. Their key strategy was to build a relatively clean training dataset from large volume of Twitter data with minimum human efforts. Their real world testing demonstrates the capability of their system to identify major life events accurately. The most parallel work that we can leverage here is the method and corpus developed by Liu et al. liu2016understanding, which is an effective supervised learning system to detect job-related tweets from individual and business accounts. To fully utilize the existing resources, we build upon the corpus by Liu et al. liu2016understanding to construct and contribute our more fine-grained corpus of job-related discourse with improvements of the classification methods.

Data and Methods

Figure FIGREF4 shows the workflow of our humans-in-the-loop framework. It has multiple iterations of human annotations and automatic machine learning predictions, followed by some linguistic heuristics, to extract job-related tweets from personal and business accounts.

Compared to the framework introduced in BIBREF16 , our improvements include: introducing a new rule-based classifier (INLINEFORM0), conducting an additional round of crowdsourcing annotations

(R4) to enrich the human labeled data, and training a classification model with enhanced performances (INLINEFORM1) which was ultimately used to label the unseen data.

Data Collection

Using the DataSift Firehose, we collected historical tweets from public accounts with geographical coordinates located in a 15-counties region surrounding a medium sized US city from July 2013 to June 2014. This one-year data set contains over 7 million geo-tagged tweets (approximately 90% written in English) from around 85,000 unique Twitter accounts. This particular locality has geographical diversity, covering both urban and rural areas and providing mixed and balanced demographics. We could apply local knowledge into the construction of our final job-related corpus, which has been approved very helpful in the later experiments.

Initial Classifier 0 \mathbf {C_0}

In order to identify probable job-related tweets which are talking about paid positions of regular employment while excluding noises (such as students discussing homework or school-related activities, or people complimenting others), we defined a simple term-matching classifier with inclusion and exclusion terms in the first step (see Table TABREF9).

Classifier INLINEFORM0 consists of two rules: the matched tweet must contain at least one word in the Include lexicon and it cannot contain any word in the Exclude lexicon. Before applying filtering rules, we pre-processed each tweet by (1) converting all words to lower cases; (2) stripping out punctuation and special characters; and (3) normalizing the tweets by mapping out-of-vocabulary phrases (such as abbreviations and acronyms) to standard phrases using a dictionary of more than 5,400 slang terms in the Internet.

This filtering yielded over 40,000 matched tweets having at least five words, referred as job-likely.

Crowdsourced Annotation R1

Our conjecture about crowdsourced annotations, based on the experiments and conclusions from BIBREF17 , is that non-expert contributors could produce comparable quality of annotations when evaluating against those gold standard annotations from experts. And it is similarly effective to use the labeled tweets with high inter-annotator agreement among multiple non-expert annotators from crowdsourcing platforms to build robust models as doing so on expert-labeled data.

We randomly chose around 2,000 job-likely tweets and split them equally into 50 subsets of 40 tweets each. In each subset, we additionally randomly duplicated five tweets in order to measure the intra-annotator agreement and consistency. We then constructed Amazon Mechanical Turk (AMT) Human Intelligence Tasks (HITs) to collect reference annotations from crowdsourcing workers. We assigned 5 crowdworkers to each HIT—this is an empirical scale for crowdsourced linguistic annotation tasks suggested by previous studies BIBREF18 , BIBREF19 . Crowdsourcing workers were required to live in the United States and had records of approval rating of 90% or better. They were instructed to read each tweet and answer following question “Is this tweet about job or employment?”: their answer Y represents job-related and N represents not job-related. Workers were allowed to work on as many distinct HITs as they liked.

We paid each worker $1.00 per HIT and gave extra bonuses to those who completed multiple HITs. We rejected workers who did not provide consistent answers to the duplicate tweets in each HIT. Before publishing the HITs to crowdsourcing workers, we consulted with Turker Nation to ensure that we treat and compensate workers fairly for their requested tasks.

Given the sensitive nature of this work, we anonymized all tweets to minimize any inadvertent disclosure of personal information (INLINEFORM0 names) or cues about an individual's online identity (URLs) before publishing tweets to crowdsourcing workers. We replaced INLINEFORM1 names with INLINEFORM2 , and recognizable URLs with INLINEFORM3 . No attempt was ever made to contact or interact with any user.

This labeling round yielded 1,297 tweets labeled with unanimous agreement among five workers, i.e. five workers gave the same label to one tweet—1,027 of these were labeled job-related, and the rest 270 were not job-related. They composed the first part of our human-annotated dataset, named as Part-1.

Training Helper Labeler 1 \mathbf {C_1}

We relied on the textual representations—a feature space of n-grams (unigrams, bigrams and trigrams)—for training. Due to the noisy nature of Twitter, where users frequently write short, informal spellings and grammars, we pre-processed input data as the following steps: (1) utilized a revised Twokenizer system which was specially trained on Twitter texts BIBREF20 to tokenize raw messages, (2) completed stemming and lemmatization using WordNet Lemmatizer BIBREF21 .

Considering the class imbalance situations in the training dataset, we selected the optimal learning parameters by grid-searching on a range of class weights for the positive (job-related) and negative (not job-related) classes, and then chose the estimator that optimized F1 score, using 10-fold cross validation.

In Part-1 set, there are 1,027 job-related and 270 not job-related tweets. To construct a balanced training set for INLINEFORM0 , we randomly chose 757 tweets outside the job-likely set (which were classified as negative by INLINEFORM1). Admittedly these additional samples do not necessarily represent the true negative tweets (not job-related) as they have not been manually checked. The noise introduced into the

framework would be handled by the next round of crowdsourced annotations.

We trained our first SVM classification model INLINEFORM0 and then used it to label the remaining data in our data pool.

Crowdsourced Annotation R2

We conducted the second round of labeling on a subset of INLINEFORM0 -predicted data to evaluate the effectiveness of the aforementioned helper INLINEFORM1 and collect more human labeled data to build a class-balanced set (for training more robust models).

After separating positive- and negative-labeled (job-related vs. not job-related) tweets, we sorted each class in descending order of their confidence scores. We then spot-checked the tweets to estimate the frequency of job-related tweets as the confidence score changes. We discovered that among the top-ranked tweets in the positive class about half, and near the separating hyperplane (i.e., where the confidence scores are near zero) almost none, are truly job-related.

We randomly selected 2,400 tweets from those in the top 80th percentile of confidence scores in positive class (Type-1). The Type-1 tweets are automatically classified as positive, but some of them may not be job-related in the ground truth. Such tweets are the ones which INLINEFORM0 fails though INLINEFORM1 is very confident about it. We also randomly selected about 800 tweets from those tweets having confidence scores closest to zero approaching from the positive side, and another 800 tweets from the negative side (Type-2). These 1,600 tweets have very low confidence scores, representing those INLINEFORM2 cannot clearly distinguish. Thus the automatic prediction results of the Type-2 tweets have a high chance being wrongly predicted. Hence, we considered both the clearer core and at the gray zone periphery of this meaningful phenomenon.

Crowdworkers again were asked to annotate this combination of Type-1 and Type-2 tweets in the same fashion as in R1. Table TABREF18 records annotation details.

Grouping Type-1 and Type-2 tweets with unanimous labels in R2 (bold columns in Table TABREF18), we had our second part of human-labeled dataset (Part-2).

Training Helper Labeler 2 \mathbf {C_2}

Combining Part-1 and Part-2 data into one training set—4,586 annotated tweets with perfect inter-annotator agreement (1748 job-related tweets and 2838 not job-related), we trained the machine labeler INLINEFORM0 similarly as how we obtained INLINEFORM1 .

Community Annotation R3

Having conducted two rounds of crowdsourced annotations, we noticed that crowdworkers could not reach consensuses on a number of tweets which were not unanimously labeled. This observation intuitively suggests that non-expert annotators inevitably have diverse types of understanding about the job topic because of its subjectivity and ambiguity. Table TABREF21 provides examples (selected from both R1 and R2) of tweets in six possible inter-annotator agreement combinations.

Two experts from the local community with prior experience in employment were actively introduced into this phase to review tweets on which crowdworkers disagreed and provided their labels. The tweets with unanimous labels in two rounds of crowdsourced annotations were not re-annotated by experts because unanimous votes are hypothesized to be reliable as experts' labels. Table TABREF22 records the numbers of tweets these two community annotators corrected.

We have our third part of human-annotated data (Part-3): tweets reviewed and corrected by the community annotators.

Training Helper Labeler 3 \mathbf {C_3}

Combining Part-3 with all unanimously labeled data from the previous rounds (Part-1 and Part-2) yielded 2,645 gold-standard-labeled job-related and 3,212 not job-related tweets. We trained INLINEFORM0 on this entire training set.

Crowdsourced Validation of 0 \mathbf {C_0}, 1 \mathbf {C_1}, 2 \mathbf {C_2} and 3 \mathbf {C_3}

These three learned labelers (INLINEFORM0 , INLINEFORM1 , and INLINEFORM2) are capable to annotate unseen tweets automatically. Their performances may vary due to the progressively increasing size of training data.

To evaluate the models in different stages uniformly—including the initial rule-based classifier INLINEFORM0 —we adopted a post-hoc evaluation procedure: We sampled 400 distinct tweets that have not been used before from the data pool labeled by INLINEFORM1 , INLINEFORM2 , INLINEFORM3 and INLINEFORM4 respectively (there is no intersection between any two sets of samples). We had these four classifiers to label this combination of 1600-samples test set. We then asked crowdsourcing workers to validate a total of 1,600 unique samples just like our settings in previous rounds of crowdsourced annotations (R1 and R2). We took the majority votes (where at least 3 out of 5 crowdsourcing workers agreed) as reference labels for these testing tweets.

Table TABREF25 displays the classification measures of the predicted labels as returned by each model

against the reference labels provided by crowdsourcing workers, and shows that INLINEFORM0 outperforms INLINEFORM1 , INLINEFORM2 and INLINEFORM3 .

Crowdsourced Annotation R4

Even though INLINEFORM0 achieves the highest performance among four, it has scope for improvement. We manually checked the tweets in the test set that were incorrectly classified as not job-related and focused on the language features we ignored in preparation for the model training. After performing some pre-processing on the tweets in false negative and true positive groups from the above testing phase, we ranked and compared their distributions of word frequencies. These two rankings reveal the differences between the two categories (false negative vs. true positive) and help us discover some signal words that were prominent in false negative group but not in true positive—if our trained models are able to recognize these features when forming the separating boundaries, the prediction false negative rates would decrease and the overall performances would further improve.

Our fourth classifier INLINEFORM0 is rule-based again and to extract more potential job-related tweets, especially those would have been misclassified by our trained models. The lexicons in INLINEFORM1 include the following signal words: career, hustle, wrk, employed, training, payday, company, coworker and agent.

We ran INLINEFORM0 on our data pool and randomly selected about 2,000 tweets that were labeled as positive by INLINEFORM1 and never used previously (i.e., not annotated, trained or tested in INLINEFORM2 , INLINEFORM3 , INLINEFORM4 , and INLINEFORM5). We published these tweets to crowdsouring workers using the same settings of R1 and R2. The tweets with unanimously agreed labels in R4 form the last part of our human-labeled dataset (Part-4).

Table TABREF27 summarizes the results from multiple crowdsourced annotation rounds (R1, R2 and R4).

Training Labeler 5 \mathbf {C_5}

Aggregating separate parts of human-labeled data (Part-1 to Part-4), we obtained an integrated training set with 2,983 job-related tweets and 3,736 not job-related tweets and trained INLINEFORM0 upon it. We tested INLINEFORM1 using the same data in crowdsourced validation phase (1,600 tested tweets) and discovered that INLINEFORM2 beats the performances of other models (Table TABREF29).

Table TABREF30 lists the top 15 features for both classes in INLINEFORM0 with their corresponding weights. Positive features (job-related) unearth expressions about personal job satisfaction (lovemyjob) and announcements of working schedules (day off, break) beyond our rules defined in INLINEFORM1 and INLINEFORM2 . Negative features (not job-related) identify phrases to comment on others' work (your work, amazing job, awesome job, nut job) though they contain “work” or “job,” and show that school- or game-themed messages (college career, play) are not classified into the job class which meets our original intention.

End-to-End Evaluation

The class distribution in the machine-labeled test data is roughly balanced, which is not the case in real-world scenarios, where not-job-related tweets are much more common than job-related ones.

We proposed an end-to-end evaluation: to what degree can our trained automatic classifiers (INLINEFORM0 , INLINEFORM1 , INLINEFORM2 and INLINEFORM3) identify job-related tweets in the real world? We introduced the estimated effective recall under the assumption that for each model, the

error rates in our test samples (1,600 tweets) are proportional to the actual error rates found in the entire one-year data set which resembles the real world. We labeled the entire data set using each classifier and defined the estimated effective recall INLINEFORM4 for each classifier as INLINEFORM5

where INLINEFORM0 is the total number of the classifier-labeled job-related tweets in the entire one-year data set, INLINEFORM1 is the total of not job-related tweets in the entire one-year data set, INLINEFORM2 is the number of classifier-labeled job-related tweets in our 1,600-sample test set, INLINEFORM3 , and INLINEFORM4 is the recall of the job class in our test set, as reported in Tables TABREF25 and TABREF29 .

Table TABREF32 shows that INLINEFORM0 can be used as a good classifier to automatically label the topic of unseen data as job-related or not.

Determining Sources of Job-Related Tweets

Through observation we noticed some patterns like:

“Panera Bread: Baker - Night (#Rochester, NY) HTTP://URL #Hospitality #VeteranJob #Job #Jobs #TweetMyJobs”

in the class of job-related tweets. Nearly every job-related tweet that contained at least one of the following hashtags: #veteranjob, #job, #jobs, #tweetmyjobs, #hiring, #retail, #realestate, #hr also had a URL embedded. We counted the tweets containing only the listed hashtags, and the tweets having both the queried hashtags and embedded URL, and summarized the statistics in Table TABREF34 . By spot checking we found such tweets always led to recruitment websites. This observation suggests that these tweets with similar “hashtags + URL” patterns originated from business agencies or companies instead of

personal accounts, because individuals by common sense are unlikely to post recruitment advertising.

This motivated a simple heuristic that appeared surprisingly effective at determining which kind of accounts each job-related tweet was posted from: if an account had more job-related tweets matching the “hashtags + URL” patterns than tweets in other topics, we labeled it a business account; otherwise it is a personal account. We validated its effectiveness using the job-related tweets sampled by the models in crowdsourced evaluations phase. It is essential to note that when crowdsourcing annotators made judgment about the type of accounts as personal or business, they were shown only one target tweet—without any contexts or posts history which our heuristics rely on.

Table TABREF35 records the performance metrics and confirms that our heuristics to determine the sources of job-related tweets (personal vs. business accounts) are consistently accurate and effective.

We used INLINEFORM0 to detect (not) job-related tweets, and applied our linguistic heuristics to further separate accounts into personal and business groups automatically.

Annotation Quality

To assess the labeling quality of multiple annotators in crowdsourced annotation rounds (R1, R2 and R4), we calculated Fleiss' kappa BIBREF22 and Krippendorff's alpha BIBREF23 measures using the online tool BIBREF24 to assess inter-annotator reliability among the five annotators of each HIT. And then we calculated the average and standard deviation of inter-annotator scores for multiple HITs per round. Table TABREF36 records the inter-annotator agreement scores in three rounds of crowdsourced annotations.

The inter-annotator agreement between the two expert annotators from local community was assessed using Cohen's kappa BIBREF26 as INLINEFORM0 which indicates empirically almost excellent. Their

joint efforts corrected more than 90% of tweets which collected divergent labels from crowdsourcing workers in R1 and R2.

We observe in Table TABREF36 that annotators in R2 achieved the highest average inter-annotator agreements and the lowest standard deviations than the other two rounds, suggesting that tweets in R2 have the highest level of confidence being related to job/employment. As shown in Figure FIGREF4 , the annotated tweets in R1 are the outputs from INLINEFORM0 , the tweets in R2 are from INLINEFORM1 , and the tweets in R4 are from INLINEFORM2 . INLINEFORM3 is a supervised SVM classifier, while both INLINEFORM4 and INLINEFORM5 are rule-based classifiers. The higher agreement scores in R2 indicate that a trained SVM classifier can provide more reliable and less noisy predictions (i.e., labeled data). Further, higher agreement scores in R1 than R4 indicates that the rules in INLINEFORM6 are not intuitive as that in INLINEFORM7 and introduce ambiguities. For example, tweets “What a career from Vince young!” and “I hope Derrick Rose plays the best game of his career tonight” both use career but convey different information: the first tweet was talking about this professional athlete's accomplishments while the second tweet was actually commenting on the game the user was watching. Hence crowdsourcing workers working on INLINEFORM8 tasks read more ambiguous tweets and solved more difficult problems than those in INLINEFORM9 tasks did. Considering that, it is not surprising that the inter-annotator agreement scores of R4 are the worst.

Dataset Description

Our dataset is available as a plain text file in JSON format. Each line represents one unique tweet with five attributes identifying the tweet id (tweet_id, a unique identification number generated by Twitter for each tweet), topics job vs. notjob labeled by human (topic_human) and machine (topic_machine), and sources personal vs. business labeled by human (source_human) and machine (source_machine). NA represents “not applicable.” An example of tweet in our corpus is shown as follows:

{

"topic_human":"NA",

"tweet_id":"409834886405832705",

"topic_machine":"job",

"source_machine":"personal",

"source_human":"NA"

}

Table TABREF37 provides the main statistics of our dataset w.r.t the topic and source labels provided by human and machine.

Conclusion

We presented the Twitter Job/Employment Corpus and our approach for extracting discourse on work from public social media. We developed and improved an effective, humans-in-the-loop active learning framework that uses human annotation and automatic predictions over multiple rounds to label automatically data as job-related or not job-related. We accurately determine whether or not Twitter accounts are personal or business-related, according to their linguistic characteristics and posts history. Our crowdsourced evaluations suggest that these labels are precise and reliable. Our classification framework could be extended to other open-domain problems that similarly lack high-quality labeled

ground truth data.