Could Twitter Help You Find the Perfect Job?

Marnie Boyer
Thinkful Data Science Bootcamp

My Experiment

Job and career changes can be stressful and uncertain. Sometimes the perfect job or company "on paper" is a nightmare in real life. Making the decision on where to pursue your career is risky and any way to make this easier would improve the likelihood of happiness and success.

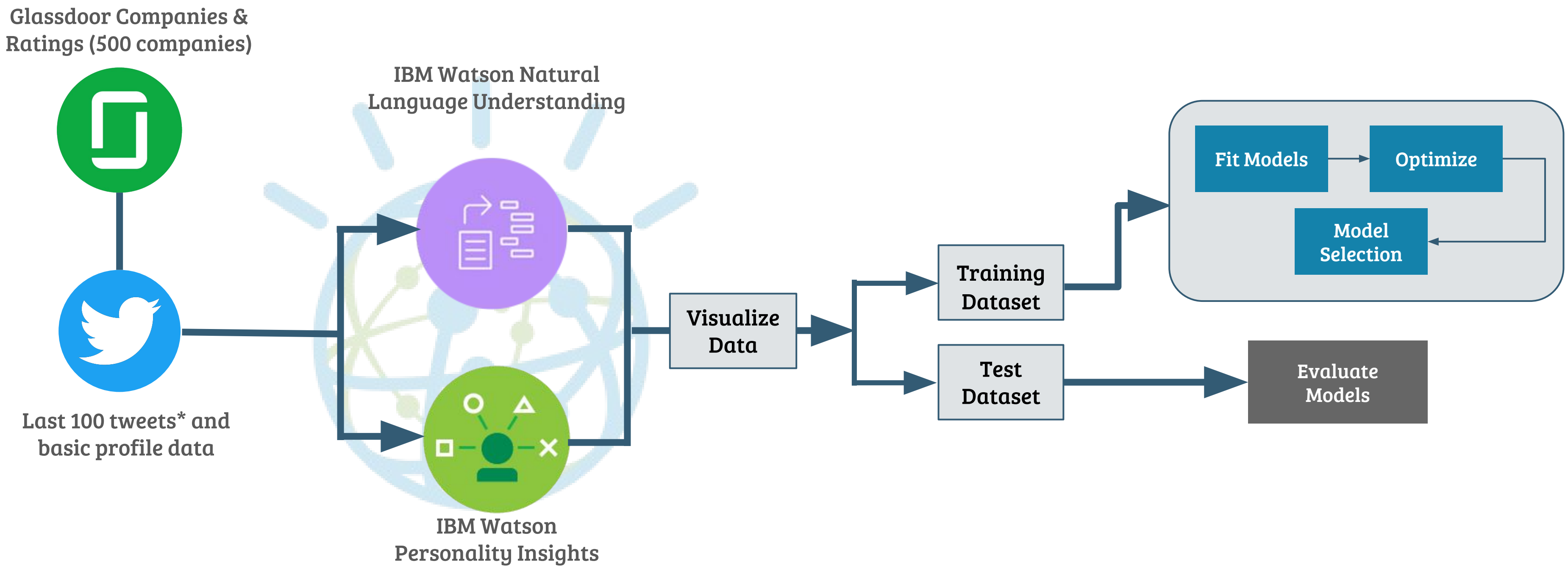
Companies have been increasingly using their social media accounts to express their opportunities and culture to attract new talent. My hypothesis is that a company's Twitter account can indicate whether it is a good company for which to work.

But how do we determine what a "good company to work for" is? Many people rely on the ratings and reviews from Glassdoor before they apply and/or accept an offer from a prospective employer, so I will use the Glassdoor categories (Dissatisfied, OK, Satisfied, Very Satisfied) to determine if this is a desirable employer.

First I examined the personality of their career and/or company Twitter accounts utilizing IBM Watson Services. Next, I developed a model to predict the Glassdoor rating category for each company.

In addition to the personality of the Twitter account, I also considered the level of effort put into and the social influence of the handle. This was determined by how long the Twitter account had been open, the number of tweets and favorites and the follower/friend counts and ratio.

Data Processing Pipeline



* Career handles scraped from company's career website

IBM Watson

By studying the authors of social posts (in this case, the employer), you can learn about what values and personality types they want to portray and gain a sense of the type of people who work there.

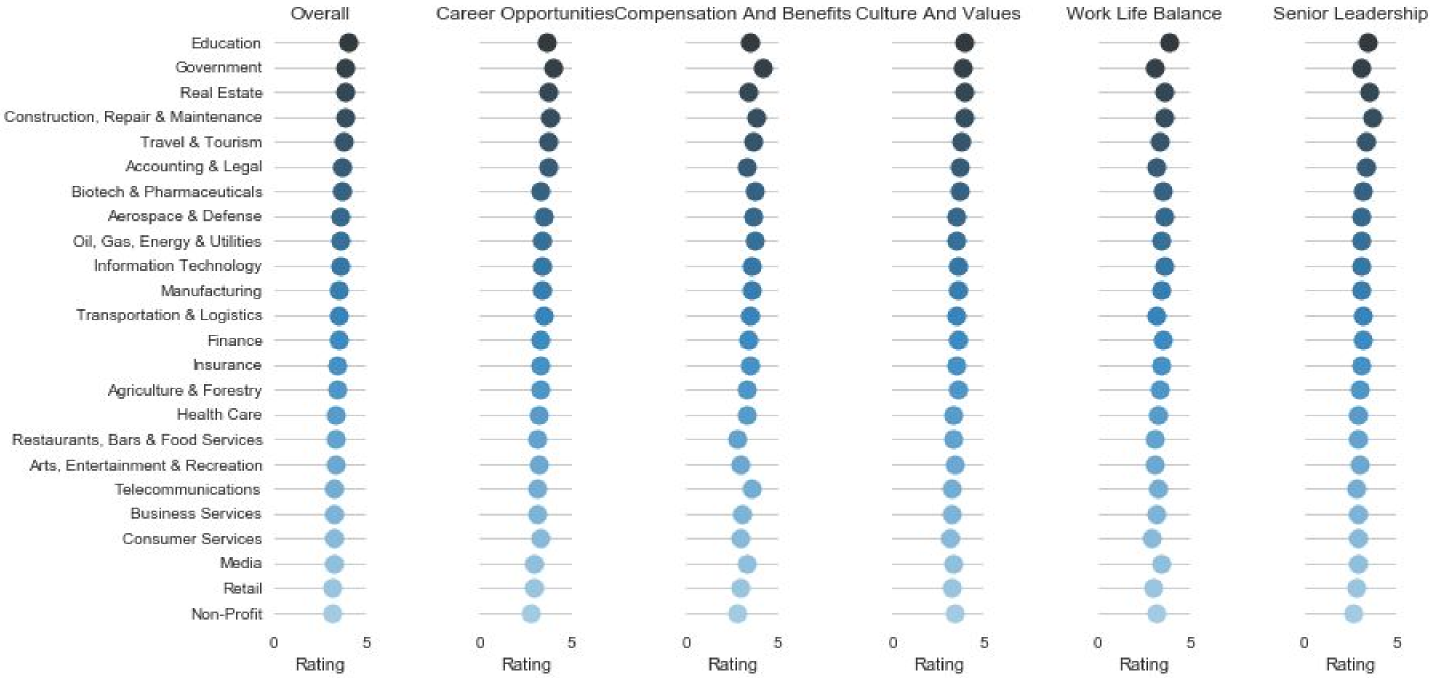
Personality Insights provides insight into how and why people think, act, and feel the way they do. PIs uses linguistic analytics to infer individuals' personality characteristics, including Big Five, Needs, and Values, from digital communications such as email, blogs, tweets, and forum posts.

Natural Language Understanding extracts metadata from text to identify concepts, entities, keywords, categories, relations and semantic roles. In my project, I used NLU to uncover the emotions (anger, disgust, fear, joy, sadness) in the employers' tweets.

Visualization of Personality Data

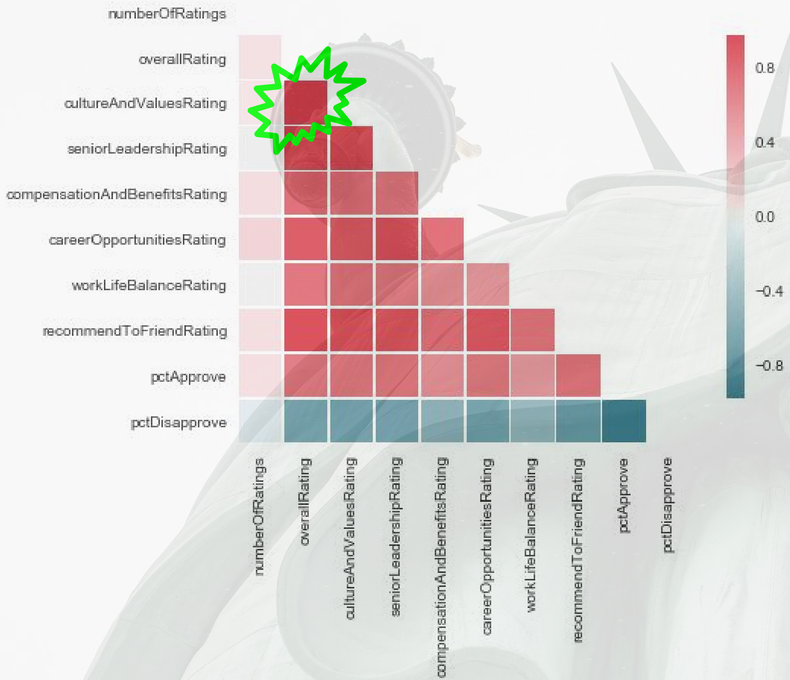
Understanding the Data

The Education industry's overall leadership is fueled by strong scores in **Culture and Values**, Work Life Balance and Senior Leadership.



Source: Glassdoor API

Satisfaction with Culture & Values is a leading indicator to overall satisfaction.



How can you determine a company's culture and values without working there?

Perhaps, like many of us, an employer's personality is demonstrated through its social presence?

Data Exploration

Top 5: # of Followers

@nike	7.1M
@intel	4.7M
@blackberry	4.5M
@spacex	4.3M
@michaelkors	3.6M

Two-thirds of employers have less than 50K followers.

Top 5: # of GD Rating

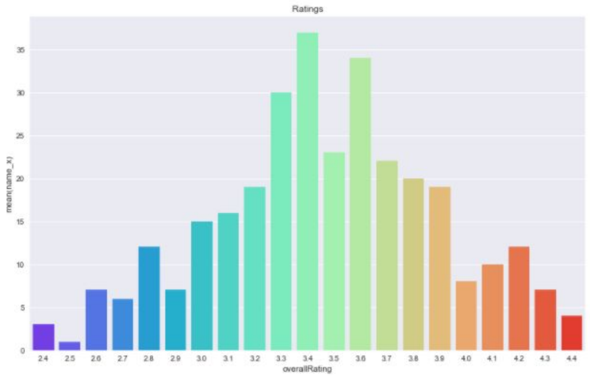
SpaceX	4.4
Insight Global	4.4
Google	4.4
Morrison HC	4.4
U of Michigan	4.3

Worst Rated Companies:

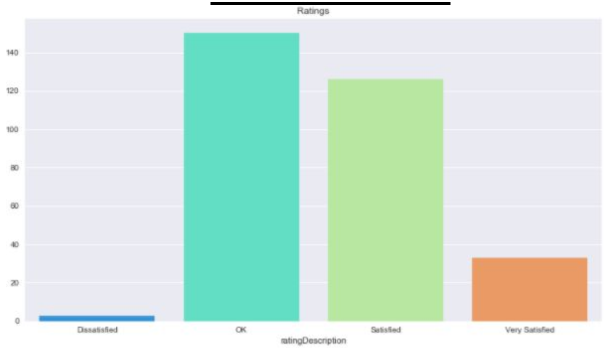
Kraft Heinz Co.	2.4
The Fresh Market	2.4
Alorica	2.4

Average # of ratings is 3,500 (range from 1K to 30K)

Rating Distribution:



Rating Description Distribution:



Data Exploration

To test the code and get more familiar with the outputs, I ran a small test and output the results to Excel. I confirmed that the output represented the desired variables (i.e., the PI and NLU variables for each company's tweets and Glassdoor ratings for each.) I ordered the data by the Glassdoor overall rating to uncover patterns in the dataset.

Perhaps no surprise, Sears has a relatively low Glassdoor rating (2.6) and it also scored lowest in Joy (a NLU variable) and among the highest on Anger. Upon reviewing the text of the tweets, they appear to be announcements of jobs available at Sears locations.

A	B	C	D	E	F	AL	AO	AQ	AR	AS	AT	AU	AV	AW	AX	A
handle	anger	disgust	fear	joy	sadness	all_text	overallRating	cultureAndValuesRating	seniorLeadershipRating	compensationAndBenefitsRating	careerOpportunitiesRating	workLifeBalanceRating	recommendToFriendRating	sectorName	pctApprove	pctD prov
@searsjobs	0.09	0.06	0.05	0.19	0.09	Assistant Store Ma... - Sears Logistics	2.6	2.5	2.2	2.3	2.4	2.8	27	Retail	15	
@cvshealthjobs	0.06	0.06	0.06	0.47	0.09	Can you recommend anyone for this #	2.8	2.7	2.4	2.7	2.8	2.6	37	Health Ca	40	
@macysjobs	0.04	0.03	0.04	0.72	0.54	Last day to help support and save up	3.2	3.2	2.8	2.8	2.9	3.1	52	Retail	66	
@walgreensjobs	0.04	0.04	0.05	0.71	0.07	Swoon! ðŸ˜Our Beauty Consultants h	3.2	3.1	2.8	3.1	3	2.9	52	Retail	50	
@walmarttoday	0.04	0.06	0.04	0.67	0.11	Hi MJ, please call our Open Door Hot	3.2	3	2.7	3.2	3.3	2.9	56	Retail	67	
@attcareers	0.09	0.06	0.07	0.50	0.08	Want to work at AT&T? We're #hi	3.4	3.2	2.9	3.8	3.2	3	60	Telecomm	67	
@attjobs	0.07	0.05	0.09	0.68	0.57	At the AT&T Foundry, diverse skil	3.4	3.2	2.9	3.8	3.2	3	60	Telecomm	67	
@lowescareers	0.06	0.06	0.06	0.61	0.12	We believe what you do away from wo	3.4	3.3	2.8	3.5	3.1	2.9	62	Retail	48	
@targetcareers	0.05	0.04	0.05	0.53	0.08	To join the Target family, we would r	3.4	3.5	2.9	3.2	3.1	3.1	64		74	
@upsjobs	0.07	0.05	0.06	0.64	0.16	Inside Sales Reps train w/ experts to b	3.5	3.3	3	3.7	3.5	3.1	66	Transport	77	
@wellsfargojobs	0.07	0.07	0.09	0.54	0.12	Our team members spent this afterno	3.5	3.5	3	3.5	3.5	3.5	67		82	
@wellsfargoworks	0.06	0.04	0.06	0.67	0.08	Is your business a functional or divisi	3.5	3.5	3	3.5	3.5	3.5	67		82	
@amazoncareers	0.09	0.08	0.08	0.63	0.53	Hereâ€™s Why Good Employees Quit	3.6	3.5	3.1	3.7	3.6	3	67	Retail	82	
@bestbuy_careers	0.06	0.05	0.05	0.70	0.10	We are currently searching for an Ass	3.7	3.8	3.2	3.4	3.3	3.3	77		89	
@starbucksjobs	0.07	0.05	0.06	0.72	0.10	Bringing the coffee journey to life thro	3.9	4	3.2	3.9	3.5	3.5	78	Restauran	86	
@microsoftjobs	0.06	0.07	0.06	0.71	0.10	Work, lunch, networkingâ€" thereâ€™	4	3.9	3.5	4.1	3.7	3.7	85	Informatio	96	

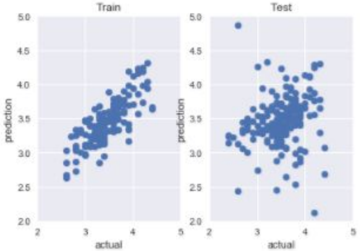
Regression Models to predict the Glassdoor rating were insufficient

Random Forest

```
In [20]: rfr = ensemble.RandomForestRegressor()

         rfr.fit(data_train,target_train)
         cross_val_score(rfr, data_train, target_train, cv=5)

Out[20]: array([-0.21812765, -0.13223685,  0.05638099,  0.12832994, -0.01014717])

In [21]: print(rfr.score(data_train,target_train))
         print(rfr.score(data_test,target_test))

0.828360814217
0.133443048442
```

LINEAR REGRESSION

```
regr = linear_model.LinearRegression()
regr.fit(data_train,target_train)
```



LASSO

```
R² for the model (train):
0.0247978488419
R² for the model (test):
-0.0420573581703
```

Other unsuccessful models attempted: Weighted & Un-weighted KNN Regression

Classification Models to predict the rating description improved results

Random Forest Classifier

```
In [252]: from sklearn import ensemble
          from sklearn.model_selection import cross_val_score

          rfc = ensemble.RandomForestClassifier()
          cross_val_score(rfc, data_train, Ctarget_train, cv=5)
          rfc.fit(data_train, Ctarget_train)
```

```
In [256]: y_pred_train = rfc.predict(data_train)
          y_pred_test = rfc.predict(data_test)


          # Display our results.
          print("Train:Number of mislabeled points out of a total {} points : {}".format(
              data_train.shape[0],
              (Ctarget_train != y_pred_train).sum()
          ))
          print("Test: Number of mislabeled points out of a total {} points : {}".format(
              data_test.shape[0],
              (Ctarget_test != y_pred_test).sum()
          ))
```

```
Train:Number of mislabeled points out of a total 134 points : 4
Test: Number of mislabeled points out of a total 137 points : 59
```

Other models attempted:
KNN Classifier: Train R^2 = 0.62; Test R^2 = 0.45
SVM failed!

Most important features

Imagination

Twitter Followers

Education

Influence Ratio

Self-discipline

Dutifulness

Artistic Interests

Conclusions & Outcomes

After many attempts at building a regression model to predict the Glassdoor Overall Rating (a number between 0.0 and 5.0), I switched to utilizing classifier models to instead predict the Overall Rating Description (represented by the classifications: Dissatisfied, OK, Satisfied, Very Satisfied).

After attempting a few classifiers, I determined that the Random Forest Classifier performed the best.

With an R^2 of 0.59 for the test data set and a 60% prediction accuracy, the model is showing promise.

Turnover is costly

Employee fit is critical. There are direct and indirect costs to replacing employees, such as:

- Cost of hiring a new person (advertising, interviewing, screening, hiring)
- Cost of onboarding and training a new person
- Lost productivity and work quality
- Lost engagement and cultural impact (others see high turnover disengage, and lose productivity)

A future application of this technique would be to develop a candidate match tool that examines a prospective employee's social handles and compares it to that of the employer. With this, candidates could select their future employers based on highest match rates and employers could evaluate fit of potential employees into its organization.

This could save millions of dollars in hiring and development costs with very little investment.

Future Ideas

- Employers could gather the social handles of their best and worst performing employees, run through the PI and NLU services and determine which personality traits are associated with the highest (and lowest) performing employees and highest (and lowest) leadership potential. (Company or department specific models)
- Determine best practices on how the best rated companies use social media
- Develop different models based on company size, and industry
- Utilize personalities to determine cultural fits for M&A activity