Using Predictive Models to Determine American Tech Workers' Likelihood to Discuss Mental Health at Work

Miriam Callahan

Introduction

With $30 billion lost annually in the US from employees taking sick days caused by work-related stress and anxiety alone, it is clear an employee's poor mental well-being at work not only has emotional implications for them but fiscal implications on the company [1]. Establishing a workplace culture which prioritizes mental health is demonstrated to decrease these effects.

The following project tested the use of various data mining algorithms to determine a technology employee's likelihood of discussing a mental health issue with their coworkers or supervisor(s) since discussing mental health issues foments and reproduces a workplace dynamic that values an employee's mental well-being and having a support system is associated with a better well-being within a work environment. It can also be used as an indicator of overall employee workplace satisfaction as those who are more willing to be transparent are likelier more satisfied and content in their position. Data was taken from a survey by the non-profit organization Open Sourcing Mental Illness from August 27th-28th, 2014 [2]. The project was made up of 3 primary phases: 1) Data Preparation, 2) Feature Selection 3) Model Development and Testing.

The Importance of Discussing Mental Health in the Workplace in Technology

With the average tech professional working 60+ hours a week and "constantly available by smartphone," it is no surprise that 57% of professionals from major companies' report feeling significant stress or burnout as a result of their jobs (See Figure 2). Nearly half of companies in the US do not provide workplace health programs for employees [3]. But in fast-paced, and stressful industries, such as technology, workplace health programs and other endeavors that promote a 'culture of health' can be particularly important as it has numerous benefits for both the company and employees.

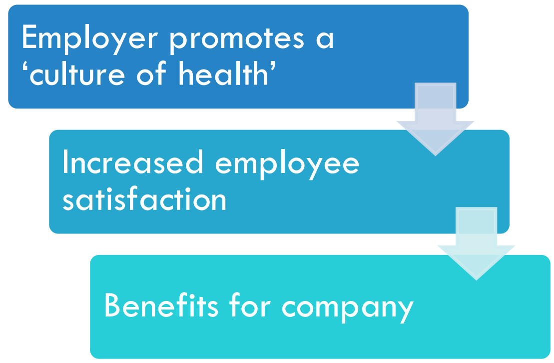


Figure #1 – A flow chart demonstrating the links between a positive workplace environment which promotes a "culture of health," employee satisfaction and company benefits

On average, an employee's productivity will increase by 12% if they are happy and for technology companies this payoff may be even higher as demonstrated by Google's example. Once the technology mogul began investing more in employee support and satisfaction such as by providing employees easy access towards mental health professionals when suffering from stress issues, more aesthetically pleasing office spaces and free lunches, productivity in the organization raised by 37% [4, 5]. This productivity not only translates into further employee satisfaction as one of their four core needs are met (physical, emotional, mental and spiritual), but also into profit for

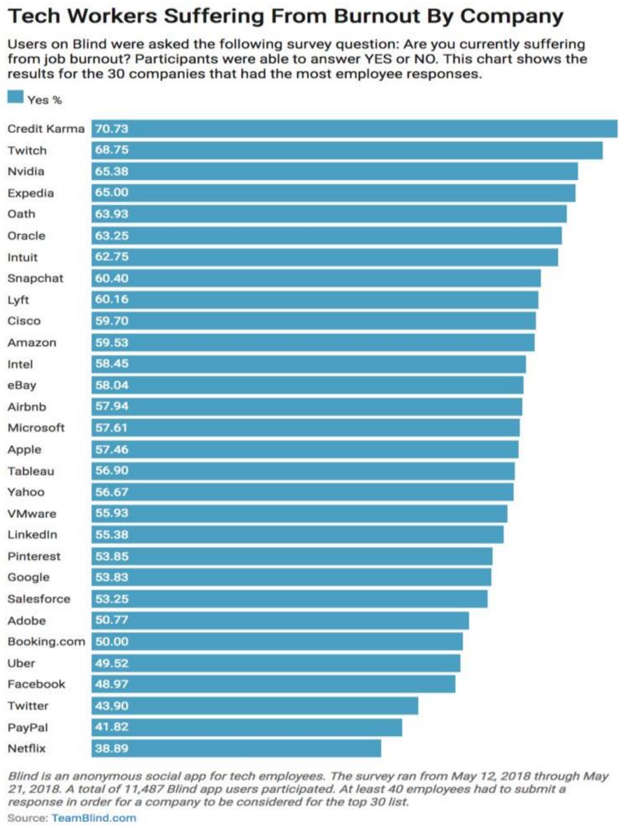


Figure #2 – A graph displaying the percentage of tech employees from major companies experiencing burnout as of May 2018 [6].

the corporation as more work is being done [7].

Outside of direct profit, employee satisfaction is also causally linked with an organization's greater financial success, particularly based on their Glassdoor ratings. Companies listed on Glassdoor with high employee satisfaction earned 1.35% more extra returns on the stock market than those who did not and research found there is a "statistically significant relationship" between employee perception of a company and a firm's future financial success [8, 9]. Generally, for every 1 star increase on a company's Glassdoor rating, their market value will jump by 7.9% [10].

Satisfied employees are also far less likely to leave their jobs. As the most important factor in employee retention, improving job satisfaction can save a company money in the long term by diverting funds that must be spent on replacing someone. On average the costs of replacing an employee are:

- 30-50% of an entry level employee's annual salary
- >150% of a mid-level employee's annual salary
- 400% of a high-level or highly-specialized employee [11].

Other than the further resources that need to be invested into obtaining a new employee, productivity is also lost as there is a standstill in whichever projects the old hire was working in.

With so many financial and emotional benefits of discussing and promoting a 'culture of health,' particularly mental health in the workplace, companies should work to improve their employees' well-being both in and out of the office for all parties' benefit.

Data Preparation

Throughout the course of the study, data was transformed three times to form different analytics base tables (ABT). The initial ABT formed by the data preparation phase was created to be tested against multiple data science algorithms to determine which features were to be deleted to improve model accuracy.

Initial Deletion of Features

The original CSV file held 26 features representing the questions in the survey. In total, there were 1,259 respondents representing 48 nations in the world. Since over half of respondents (746) were from the United States and many nations were left unrepresented, those who did not select their country as the US were omitted and the 'country' feature itself deleted. Similarly, since data collection took place over a two-day period, the timestamp feature was also deleted as time would not be an important factor in data analysis.

The last feature removed upon initial investigation was the ‘comments’ feature. Many respondents left this field blank and those who did not gave information which was too varied to be of any use.

Instance Deletion, Gender Normalization and Underrepresentation

Since certain respondents put in ages which did not exist (like -1726) or someone would be unlikely to work at (like 5), they were removed from the dataset. One respondent had put down their age as -29, but this was assumed to be a typo so the value 29 was used. Others put in genders and had responses which suggested they were not taking the survey seriously. Hence, respondent #388 who put their gender as “Nah” was also omitted.

Participants also had the option to type in their gender, resulting in there being 49 different values within the gender feature. To make this information more generalizable, the 49 different classifications were normalized to form 4 new gender variables: “Cis Male,” “Cis Female,” “Non-binary” and “Trans Female.” Distinctions between cisgender and transgender individuals were made since being transgender may significantly affect the mental health of those individuals. The value “Trans Male” would have been used would have been used as well if any of the values corresponded to that classification but none did. As a result, the data did not represent any trans men whatsoever.

Trans women and non-binary individuals were also disproportionately represented in the dataset, making up less than .012% of participants. As of 2016 however, adults who identify as transgender only make up 0.6% of the US population so this representation is valid in comparison to the general population [12]. The same occurred with cisgender women, who made up 24.13% of the sample population but similarly only make up approximately 20% of the US tech workforce [13].

Some states were also left unrepresented in the dataset: Alaska, Arkansas, Delaware, Hawaii, Montana and North Dakota. With no data from these states, no predictions could be made from the data on the likelihood that employees would discuss mental health in the office. As some of the least populous states in the nation however, little concern was given to finding some sort of representation for these states within the dataset as few workers in the technology industry would be likely to operate there anyways. Figure 3 illustrates the features left in the ABT after data preparation.

Feature Selection

Five algorithms (One R, Naïve Bayes, Repeated Incremental Pruning to Produce Error Reduction [RIPPER], C4.5 and Random Forest) were used on the version of the ABT shown in Figure 2 to determine which features were to be deleted to form the final analytics base tables for both the coworker target feature and the supervisor target feature. Each algorithm was tested on a 60/40 split of data for the training and test sets to determine feature importance/ selection. Features which were used in the model were considered important or with a GINI index greater than 10 when using the Random Forest approach.

NOTE: All GINI indexes used in this paper have been rounded to 2 decimal places for ease of readability.

Feature Selection with Coworker as a Target Feature

Of the five models produced for coworker feature deletion, Naïve Bayes had the worst

Feature	**Associated Values**
Age	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 60, 62, 65, 72
Gender	Cis Female, Cis Male, Non-binary, Trans Female
self_employed: Are you self-employed?	Yes, No
family_history: Do you have a family history of mental illness?	Yes, No
treatment: Have you sought treatment for a mental health condition?	Yes, No
work_interfere: If you have a mental health condition, do you feel that it interferes with your work?	Never, Rarely, Sometimes, Often
no_employees: How many employees does your company or organization have?	1-5, 6-25, 26-100, 100-500, 500-1000, More than 1000
remote_work: Do you work remotely (outside of an office) at least 50% of the time?	Yes, No
tech_company: Is your employer primarily a tech company/organization?	Yes, No
benefits: Does your employer provide mental health benefits?	Yes, Don't know, No
care_options: Do you know the options for mental health care your employer provides?	Yes, Not sure, No
wellness_program: Has your employer ever discussed mental health as part of an employee wellness program?	Yes, Don't know, No
seek_help: Does your employer provide resources to learn more about mental health issues and how to seek help?	Yes, Don't know, No
anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?	Yes, Don't know, No
leave: How easy is it for you to take medical leave for a mental health condition?	Very easy, Somewhat easy, Don't know, Somewhat difficult, Very difficult
mental_health_consequence: Do you think that discussing a mental health issue with your employer would have negative consequences?	Yes, Maybe, No
phys_health_consequence: Do you think that discussing a physical health issue with your employer would have negative consequences?	Yes, Maybe, No
coworkers: Would you be willing to discuss a mental health issue with your coworkers?	Yes, Some of them, No
supervisor: Would you be willing to discuss a mental health issue with your direct supervisor(s)?	Yes, Some of them, No
mental_health_interview: Would you bring up a mental health issue with a potential employer in an interview?	Yes, Maybe, No
phys_health_interview: Would you bring up a physical health issue with a potential employer in an interview?	Yes, Maybe, No
mental_vs_physical: Do you feel that your employer takes mental health as seriously as physical health?	Yes, Don't know, No
obs_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?	Yes, No

Figure #3 – The features created in the ABT after data preparation and their associated values with survey question in parentheses. The target features and their values are highlighted.

accuracy (41.81%) but produced a few attributes that were deemed important by other models (namely Random Forest whose GINI index values were highlighted the importance of certain features) such as self_employed and obs_consequences. The second worst model was One R, producing an accuracy of 53.85% based on the production of 39 rules about the state feature. Due to One R's low accuracy and large number of rules, state was considered for removal as an attribute however it was given a high GINI index by the Random Forest model (22.33) which had a high accuracy, so it was ultimately kept in the dataset.

Attributes that were deemed unimportant by all models – Gender, wellness_program, seek_help, leave, phys_health_consequence and mental_vs_physical – were immediately removed from the dataset. However, a seventh feature, tech_company, was also removed. Although the feature was deemed important by Naïve Bayes, it had one of the lowest GINI indexes when run through the Random Forest algorithm (approximately 3.5). Due to Naïve Bayes low accuracy and tech_company's low GINI value, it was removed.

Overall, the most important features appeared to be mental_health_consequence and mental_health_interview. Both were included in the rules/nodes the 2 models with the highest accuracy: RIPPER (73.67%) and C4.5 (75.63%). Mental_health_consequence also had a GINI index of 18.26 in the Random Forest model which had 66.89% accuracy and mental_health_interview almost met the threshold of importance as well with a GINI index of 8.2.

The final analytics base table developed for the coworker target feature included Age, state, self_employed, family_history, treatment, work_interfere, no_employees, remote_work, benefits, care_options, anonymity, mental_health_consequence, mental_health_interview, phys_health_interview, obs_consequence and coworkers.

Feature Selection with Supervisor as a Target Feature

Feature deletion for the supervisor dataset underwent the same process as for the coworkers' data; the data was split 60/40 and ran through five different modeling algorithms to determine which descriptive features were most important in classification.

Only one attribute was deleted from the supervisor data set: benefits. While this left the supervisor ABT significantly larger than that of the coworker data, most features were deemed important by at least 2 models in the dataset. Additionally, none of the models stuck out as drastically worse compared to the others. The RIPPER model held the lowest accuracy with 54.62% but all the models yielded similar accuracies within a range from 54 – 62%. As a result, all the features the models deemed important were taken to be important in when creating models based on the final ABT. The final supervisor ABT held all features demonstrated in Figure 3 except for the benefits attribute.

Model Development and Testing: Coworker Dataset

Since there were only 746 instances in the dataset which is not enough to split it well and form models based on appropriate amounts of data, 10-fold cross validation was performed for every algorithm to determine its accuracy. Data was divided into 10 approximately equal folds with about 75 instances in each fold and run through the model the same number of times. During each run through, one fold was removed to be used as test data with the remaining 90% as training data. The average accuracy of these models was taken to be the accuracy of the final 11th model, which was based the full ABT.

Of the four models created, RIPPER and C4.5 heralded the highest accuracy, with 63.61% and 63.60% respectively. Since these models held the highest accuracies in the initial test run, it comes as no surprise they held the highest accuracies in the final round of model creation. Nevertheless, their accuracies significantly decreased in comparison to when the omitted features were present. The Naïve Bayes model improved the most, its accuracy rising from 41.81% to 62.15%. This is likely due to the feature deletion as the fewer attributes yielded less noise in the data to confuse the algorithm.

Best Coworker Model: RIPPER

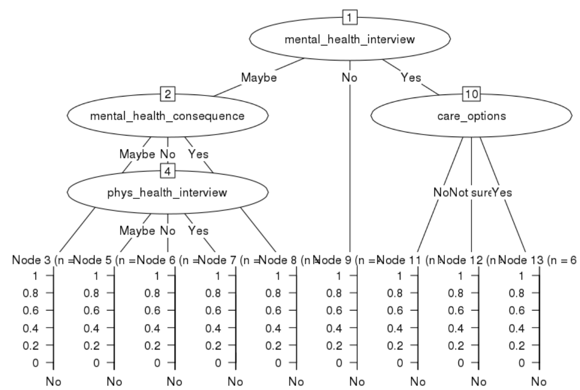
```
(mental_health_consequence = No) and (mental_health_interview = Maybe) => coworkers=Yes (33.0/16.0)
(Age >= 32) and (Age <= 33) and (phys_health_interview = Yes) => coworkers=Yes (9.0/2.0)
(mental_health_consequence = Yes) and (no_employees = More than 1000) and (obs_consequence = No) => coworkers=No (25.0/4.0)
(Age >= 43) and (treatment = No) and (Age <= 48) => coworkers=No (5.0/0.0)
 => coworkers=Some of them (285.0/72.0)

(mental_health_consequence = No) and (mental_health_interview = Maybe) and (treatment = Yes) => coworkers=Yes (29.0/10.0)
(mental_health_consequence = No) and (Age >= 39) and (benefits = Yes) => coworkers=Yes (18.0/6.0)
(mental_health_consequence = Yes) and (no_employees = More than 1000) => coworkers=No (55.0/21.0)
(mental_health_consequence = Yes) and (remote_work = No) and (Age >= 33) => coworkers=No (24.0/10.0)
 => coworkers=Some of them (458.0/134.0)
```

Figure #4 – A comparison of the RIPPER models before feature deletion (above) and after (below).

Both RIPPER models contained 5 rules and used mental_health_consequence and mental_health_interview as antecedents which the most important descriptive features were. Of the variables used, none were deleted between datasets so the significant shift in rules makes little sense. Perhaps there were hidden relationships in the data between deleted features and those used by the second model which explain why those variables were chosen for rule creation. While there is no definitive reason for why the rules changed, the accuracy reduction can be explained by this difference in rules.

Best Coworker Model: C4.5




```
mental_health_interview = Maybe
|   mental_health_consequence = Maybe: Some of them (21.0/3.0)
|   mental_health_consequence = No
|   |   phys_health_interview = Maybe: Some of them (41.0/18.0)
|   |   phys_health_interview = No: Yes (1.0)
|   |   phys_health_interview = Yes: Yes (9.0/2.0)
|   mental_health_consequence = Yes: Some of them (1.0)
mental_health_interview = No: Some of them (499.0/177.0)
mental_health_interview = Yes
|   care_options = No: Some of them (3.0)
|   care_options = Not sure: Some of them (3.0/1.0)
|   care_options = Yes: Yes (6.0)

Number of Leaves  :     9

Size of the tree :      13
```

Figure #5 – The final C4.5 coworker model produced in tree and text form.

The decision tree produced by the C4.5 algorithm had 9 leaves and 13 nodes, reducing the size of the original tree developed with 21 leaves and 32 nodes. Like the RIPPER model, its accuracy was significantly reduced from 75.63% to 63.60%. While the complexity was reduced in the new tree, none of the variables used in between trees were omitted in the dataset leading to a similar conundrum from what occurred in the RIPPER situation.

Between the two final models, mental_health_interview and mental_health_consequences were the only two variables used in both which highlights their importance. Perhaps further exploration should create models based solely on these two variables.

Although they have nearly the same accuracy, the RIPPER model is more likely to be used in real world applications in comparison to the C4.5 model which uses fewer variables and rules. Since it is simpler, the RIPPER model uses less computational power and is easier to understand. However, the decision tree is not much larger than the rules produced by RIPPER and has the added benefit that it uses fewer attributes.

Model Development and Testing: Supervisor Dataset

The supervisor data underwent the same processing as the coworker data. Like the initial models, the models produced after feature deletion yielded similar accuracies except they were in a smaller range (55 – 58%). Since only one attribute was deleted, much of the potential noise in the data remained which may be responsible for the lack of noteworthy changes in accuracy. The pattern of feature deletion

Best Supervisor Model: C4.5

Like the best coworker models, accuracy decreased between the original C4.5 model and the one based on less data albeit much less (from 61.91% to 58.16%). Although the accuracy went down slightly, the complexity of the tree produced greatly increased. The original tree held 73 leaves and 90 nodes; the later one held 94 leaves and 120 nodes. The tree also appeared to be quite similar, initially splitting on mental_health_consequence, then obs_consequence, state and mental_health_interview. The second splits and onward only occurred when mental_health_consequence equaled 'Maybe.' When the predictive feature equaled 'Yes,' the supervisor values was predicted as 'No' and vice versa. This is expected since if someone believes there will be no repercussions for discussing a mental health issue with their employer, they would be more willing to open to their supervisor if a situation arises.

As demonstrated by Figure 6, two of the rules directly created by the One R model are the same as the latter branches of the tree further corroborating the importance of mental_health_consequence as an attribute and the robustness of the rules/branches of

- If mental_health_consequence = Yes => supervisor = No
- If mental_health consequence = No => supervisor = Yes

Supervisor One R Model: A Potential Frontrunner

```
If mental_health_consequence = Maybe then supervisor = Some of them
If mental_health_consequence = No    then supervisor = Yes
If mental_health_consequence = Yes   then supervisor = No
```

Figure #6 – Rules produced by the final One R model for the supervisor dataset

With the second highest accuracy of 56.63% and only 3 rules, the One R model appears to be the best fit for deployment. Although it is unlikely it would ever be deployed due to its relatively low accuracy, the drastically lower complexity of the model makes it more appealing for real world practice as the curse of dimensionality is reduced dramatically. By only using a singular feature for prediction as well, the data set could also be reduced in size. While this is not an important consideration now as the data set is tiny, if it increases in size this could be a great advantage.

Conclusions

An individual's perception of the likelihood discussing a mental health issue in the workplace would lead to repercussions is the best indicator of how likely they are to discuss an issue like this with coworkers or a supervisor. No conclusive evidence was found suggesting that there was one "best" algorithm to use to test if someone would discuss mental health issues with coworkers or a supervisor within this data set but Naïve Bayes did not appear to be particularly helpful. In order for any of these models to be useful for corporate application, their overall accuracies must significantly increase.

ALGORITHM - COWORKERS	BEFORE	AFTER
One R	53.85%	57.45%
Naïve Bayes	41.81%	62.15%
RIPPER	73.67%	63.61%
C4.5	75.63%	63.60%

ALGORITHM - SUPERVISOR	BEFORE	AFTER
One R	57.53%	56.63%
Naïve Bayes	57.86%	56.17%
RIPPER	54.62%	55.11%
C4.5	61.91%	58.16%

Figure #7 – Charts comparing the accuracies before and after feature deletion of all model types rounded to 2 decimal places.

Within both datasets, there are two primary correlations between feature deletion and the development of the model with the highest accuracy:

1. As features are deleted between models, model accuracy goes down between 3.75 – 12.03%.
2. As features are deleted, the complexity of the model increases since more variables are used.

One could argue the accuracy decreases at a steeper rate within the data when more features are deleted, however different combinations of deleted features would need to be used to prove this idea.

Future Work

Since this project was conducted using 2014 data, there is a strong possibility that the model will be stale if new data is introduced. Open Source Mental Illness also releases a similar version of this survey annually and provides the data for free online to use so using more contemporary data would likely provide more fruitful results that could be deployed by companies. Other algorithms could also be used to develop a better model for deployment. Most of the models produced in this project had accuracies ranging from 56 – 63% accuracy, which is slightly better than guessing. In a real-world scenario, these models would likely be unable to provide much insight to companies as a marker of their workplace culture and to determine ways to improve it.

Further examination should also be put into why feature selection had an adverse effect on model accuracy and whether using different combinations of features would improve it. Perhaps a different algorithm (such as a greedy search) would be useful for determining feature deletion.

References

[1] "The high price of workplace stress", *Harvard Gazette*. [Online]. Available: https://news.harvard.edu/gazette/story/2

016/07/the-high-price-of-workplace-stress/.

[2] "Mental Health in Tech Survey - dataset by quanticdata", *Data.world*, 2017. [Online]. Available: https://data.world/quanticdata/mental-health-in-tech-survey.

[3] M. Ward, "A brief history of the 8-hour workday, which changed how Americans work", *CNBC*, 2017. [Online]. Available: https://www.cnbc.com/2017/05/03/how-the-8-hour-workday-changed-how-americans-work.html.

[4] A. Oswald, E. Proto and D. Sgroi, "Happiness and Productivity", *Journal of Labor Economics*, vol. 33, no. 4, pp. 789-822, 2015.

[5] K. Scott, "EXCLUSIVE: Google uses data to inform mental health approach", *Employee Benefits*, 2018. [Online]. Available: https://www.employeebenefits.co.uk/exclusive-google-mental-health-approach/.

[6] K. Fagan, "Employee Burnout Is Real. A Survey of More Than 11,000 Tech Workers Reveal Where It's Worst", *Inc.com*. [Online]. Available: https://www.inc.com/business-insider/employee-burnout-survey-blind-app-tech-workers.html.

[7] T. Schwartz and C. Porath, "Opinion | Why You Hate Work", *Nytimes.com*, 2014. [Online]. Available: https://www.nytimes.com/2014/06/01/opinion/sunday/why-you-hate-work.html?_r=0.

[8] T. Green, R. Huang, Q. Wen and D. Zhou, "Crowdsourced Employer Reviews and Stock Returns", *SSRN Electronic Journal*, 2018. Available: 10.2139/ssrn.3002707.

[9] A. Moniz, "Inferring Employeess Social Media Perceptions of Corporate Culture and the Link to Firm Value", *SSRN Electronic Journal*, 2017. Available: 10.2139/ssrn.2768091.

[10] M. Huang, P. Li, F. Meschke and J. Guthrie, "Family firms, employee satisfaction, and corporate performance", *Journal of Corporate Finance*, vol. 34, pp. 108-127, 2015. Available: 10.1016/j.jcorpfin.2015.08.002.

[11] K. Borysenko, K. Borysenko and A. Archive, "What Was Management Thinking? The High Cost of Employee Turnover", *TLNT*, 2015. [Online]. Available: https://www.tlnt.com/what-was-leadership-thinking-the-shockingly-high-cost-of-employee-turnover/#.

[12] The Williams Institute, "How Many Adults Identify as Transgender in the United States?", University of California: Los Angeles School of Law, 2016.

[13] "National Center for Women & Information Technology", *National Center for Women & Information Technology*. [Online]. Available: https://www.ncwit.org/.