SMU Data Science Review

Volume 3 | Number 2 Article 16

2020

Topic Modeling to Understand Technology Talent

Chad Madding
Southern Methodist University, cmadding@smu.edu

Allen Ansari
aansari@smu.edu

Chris Ballenger
cballenger@smu.edu

Aswini Thota
aswininathroy.thota@fmr.com

Topic Modeling to Understand Technology Talent

Allen Ansari[1], Chris Ballenger[1], Chad Madding[1], and Aswini Thota[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {aansari,cballenger,cmadding}@smu.edu
[2] Data Scientist, Director, Data Analytics and Insights, Fidelity Investments, Westlake TX 76262 USA {aswininathroy.thota}@fmr.com

Abstract. Attracting technology talent in today's hiring climate is more complicated than ever. Recruiting for technology talent in non-technology industries is even more challenging. This intense hiring landscape is motivating companies to attract the right talent and create a culture that can retain and grow that talent. In this paper, we developed algorithms and present insights that use data provided in reviews to glean information employers can use to address or even change their priorities to meet the demands of an ever-changing job market. The core of our research is to investigate and attribute the role of company reviews in explaining the critical dimensions through which employees perceive their job. To provide more in-depth and targeted insights, we limit our focus to Technology related job reviews. Our contributions include building an IT Professional profile that can help create an edge for recruiters. We achieved our research by conducting a comprehensive topic modeling on employee reviews to detect aspects that define technology workers. Three unique topics not related to Indeed were discovered from our models: *Learning & Development*, *Technical Skills* ,and *IT Support*. Eight additional topics show a more detailed analysis related to Indeed's topics. We feel the information provided in our paper is a tool that HR and recruiters can use to attract talent.

Keywords: text mining · people analytic · topic modeling · war for talent · hr analytic · nlp · lda · unsupervised neural network · employee reviews · Indeed · technology talent

1 Introduction

The Society for Human Resource Management shows a trend where companies find it challenging to recruit and retain workers. The report states that "HR professionals from all industries are experiencing a highly competitive market for talent, with recruiting difficulty reaching levels not seen in years."[16]

Where generations in the past were motivated by perks, recognition, and promotions, we are now in a hiring era where talent now has the opportunity to sell themselves to the highest bidder. Unlike the past two generations, loyalty is no longer the top priority. Culture is now essential to today's younger workforce[2].

This generation has no problem jumping from job to job[1]. As companies are working to attract and retain this talent, we feel one needs to be aware of what is motivating them to choose one opening over another. Our project looked at ways to use data provided in reviews to glean information employers can use to address or even change their priorities to meet the demands of an ever-changing job market.

In today's media, we see reviews on almost every website. A person can review virtually every point in their life. With this seeming overabundance of reviews, does this data hold any amount of information, or is this just rambling? We believe there are several reasons to focus on reviews. Reviews allow a company to see themselves from the outside. Reviews from potential or new employees allow the company to step outside itself and see the business from an outsider's view. What others say on websites allows those within the company to review and evaluate their way of business. Reviews allow them to see if their view of the business aligns with those outside the company.

We live in a culture that is relying more on survey options before making job-related decisions. A recent article noted that 90% of people trust and use reviews before continuing their interactions with a company[1]. Companies recognizing this can build a bond with those looking for information about their business. Google states that reviews can provide valuable feedback about a company, and commenting on reviews helps build trust with an audience[2]. Reviews give companies a way to measure customer and employee satisfaction. Negative reviews are just as critical, in that they provide an opportunity to connect with the reviewer and remedy their problems.

Review sites allow for a more level playing field for both large and small companies. Sites like Glassdoor and Indeed allow for reviews from companies of all sizes. This environment allows for smaller companies to be as much of a potential employer as larger ones. Companies, large and small, can gain insight from reviews in drawing talent.

Visitors to Indeed invest time in writing reviews, both positive and negative. With this type of energy provided in their reviews, we are looking to draw insight into the topics we will use throughout this project. Glassdoor states that 86% of users will likely research company reviews and ratings before applying[3], providing relevancy to the data collected.

Online reviews allow for a clear view of what is said about the company and the jobs they offer. Reviews assist a company in not losing focus on how the outside world views its brand. Surveys also allow for feedback but do not seem as useful in that companies can ask guided off topical questions[18] to mine specific data, whereas reviews are more open-ended.

Reviews provide a connection between the employer and those reviewing the company. Providing data to the employers will allow them to respond to those reviews to attract the talent needed at the company. Providing information back

[1] https://www.brightlocal.com/research/local-consumer-review-survey/
[2] https://support.google.com/business/answer/3474050
[3] https://www.glassdoor.com/about-us/

to the employer will allow for a reframing of their needs so that open positions show up in relevant searches.

Indeed has categorized reviews by Work-Life Balance, Pay & Benefits, Job Security & Advancement, Management, and Culture[4]. Each category is universally accepted as areas candidates research. Indeed allows users to rate each category [Fig. 6], but does not provide any sentiment analysis on reviews itself. Without clear insight into why one rating was chosen over another, recruiters may find it hard to grasp the most important topics these employees want without a detailed outlook from reviews.

Previous research correlates employee satisfaction to company performance[6,11]. The research performed used predefined culture and employee satisfaction topics to build their model. Luo et al. used word frequency from company reviews to determine if there is an impact on performance[11]. The research was able to show that employee satisfaction is different between industries. Another technique to mine employee reviews is to use Latent Dirichlet Allocation (LDA)[3], an unsupervised model. An unsupervised model can identify new topics, not previously determined through canonical research. Jung et al. used LDA to learn new topics that were not previously known, such as Software Development and Project Management[9]. Their research evaluated all employee reviews. Thus, the results were generalized and did not match each employee's satisfaction[17].

Whereas past research focused on larger groups of people and more guided surveys, we will focus on the specific job family of Technology. In doing this, we feel our results will have an advantage by narrowing the focus. With the high demand for technology specialists, companies from all sectors will benefit from our research results.

We focused on finding information for a specific type of talent. We believe that researching a specific job family will allow others to take this information and tailor our findings to attract talent to fill their specific need. We hope that this strategy will be a framework for other researchers to examine other targeted job families.

There are more software developer openings in every region of the country than qualified people to fill them. The search to find this talent can be extremely challenging[5]. We feel that information technology roles will provide the most return from our research.

The analysis we performed was a macro-level topic modeling on entire reviews and a micro-level on individual sentences. Our results concluded three new topics not previously defined by Indeed and a more granular level of topics that are related to the five topics defined by Indeed.

In the remaining paper, we will review the contribution of our research. First, we provide an overview of the demand for technology talent (Section 2). Sections 3, 4, and 5 include an overview of Indeed, the data we collected, and what models we used for topic modeling. Our topic model includes a statistically-

[4] https://www.indeed.com/

[5] https://www.arcgis.com/apps/MapJournal/index.html?appid=b1c59eaadfd945a68a59724a59dbf7b1

based model and a state of the art neural network design. Following the overview of topic modeling, we show how we implemented the models (Section 6) and results (Section 7). We will wrap up our research with a discussion about ethics (Section 8) and concluding our research (Section 9).

2 War for Information Technology Talent

We selected technology job functions as our targeted research for its large, highly skilled job pool. As more businesses strive for a competitive edge, the demand for technology job functions increases. As of 2019, Software Developers are the most in-demand job[6]. Given the demand, candidates are now in control of their destiny and entertaining multiple job offers[20]. Our research is to use the same tools as employees in order to learn the top topics.

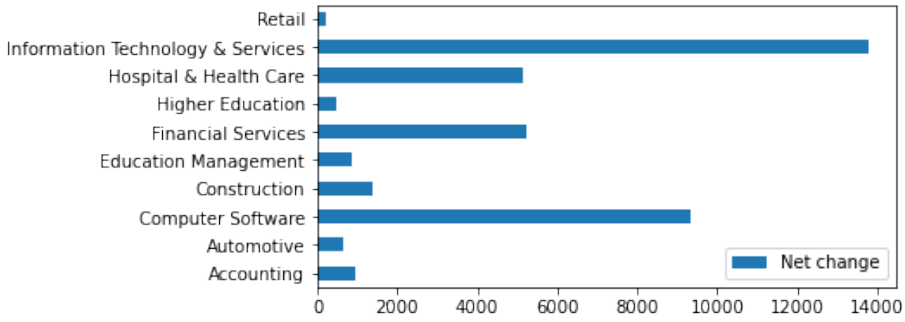

Fig. 1. Net change of IT professionals by industry

The trend for more technology in all industries makes non-traditional industries compete for the same talent. A higher number of professionals are leaving the Information Technology & Services industry and Computer Software(Fig [1]). Higher Education and Retail see some of the most significant numbers of employees from IT(Fig [2]). Due to the IT movement trend in new industries, we decided to include several industries in our data collection.

Recruiters need actionable insight to be useful in having an edge in the war for talent. We felt that time to fill, the time from posting a job to the time an offer is accepted[7], was an essential metric to consider during this research. With time to fill ranging anywhere from 51 to 57 days[8], HR professionals, business managers, and recruiters need to utilize every tool available to optimize that

[6] `https://www.cnbc.com/2019/01/24/here-are-the-most-in-demand-jobs-for-2019.html`

[7] `https://resources.workable.com/tutorial/recruiting-kpis`

[8] `https://resources.workable.com/stories-and-insights/time-to-hire-industry`

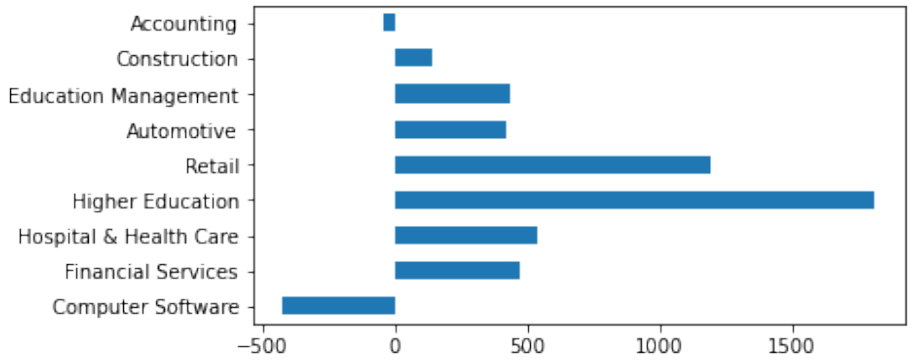

Fig. 2. Count of IT Professionals by Industry, who moved into non-traditional IT Industries from Information Technology & Services

time. For example, if time to fill runs too long then the company risks losing top talent or time to fill runs to short then the company could end up not screening property, thus making a bad hire[9]. We provide these topics to be used in the recruiting process to assist in optimizing time to fill.

3 Company Reviews

3.1 Indeed

Indeed is a website that allows employees to write reviews of companies. Indeed is the number one job site in the world, according to comScore's Total Visits in March 2018[10], with over 250 million unique visitors every month[11]. Their site has millions of personalized jobs, salary information, company reviews, and interview questions. We choose Indeed over other sites in this field because of its amount of available data and its reputation as an industry-standard company review site[12].

> "Indeed strives to put job seekers first, giving them free access to search for jobs, post resumes, and research companies. Every day, we connect millions of people to new opportunities."[13].

[9] https://resources.workable.com/stories-and-insights/time-to-hire-industry
[10] https://www.comscore.com/
[11] https://analytics.google.com/
[12] https://support.google.com/business/answer/3474050
[13] https://www.indeed.com/about

3.2 Creating a Review

When creating a review, Indeed starts by asking survey questions on a scale of 1-5[Fig. 3]. The first section ends with two multi-selection survey questions[Fig. 4].

Fig. 3. 1-5 survey question

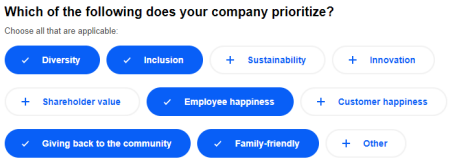

Fig. 4. Multi-selection survey question

The second page of their review section contains a 150 character minimum, free form text-box[Fig. 5] with an optional pro/con section and a star rating section[Fig. 6].

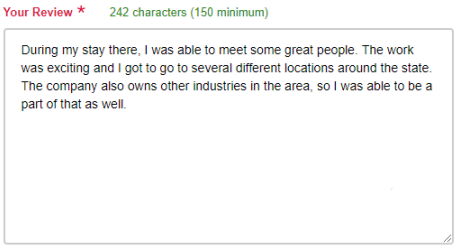

Fig. 5. 150 character minimum text-box

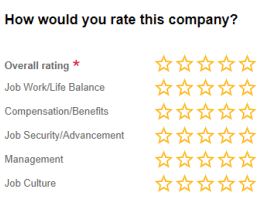

Fig. 6. Indeed's Star Rating

Indeed differs from Glassdoor in that they require users to write in each section for pros, cons, compensation, and benefits. We believe this could direct users to discuss specific topics and thus create bias in our results. Another difference is that Indeed reviews show the number of people who think one review is helpful or not helpful; on the other hand, Glassdoor just shows the number of people who think one review is helpful. The final section in a review is entitled, "Tell us about you. "This section contains the job titles we used in this research. The entry is a guided drop-down selection that Indeed has on record, but the form will accept any answer[Fig. 7]. The methods used to collect the titles will be discussed in the "Data Collected "section. The "Tell us about you "section is also where users fill out the location of the job being reviewed. The location box is also in the form of a guided selection. Users are shown selections as the location is typed, but the form will accept any entry. After a review is submitted, it must first be approved before the review will be posted[Fig. 9].

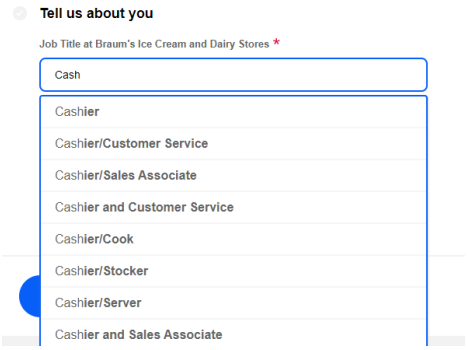

Fig. 7. Job Title from Indeed(screenshot)

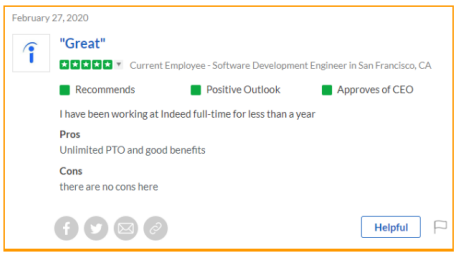

Fig. 8. Indeed's Pro/Con

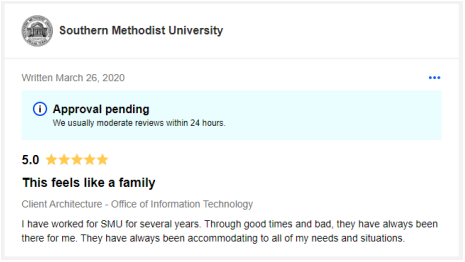

Fig. 9. Pending Approval

4 Data Collection

LinkedIn Insight[14] provided us the top 100 companies with the highest number of IT professionals within their industry. We selected the most popular industries as of 2019[15]. We built a web scrapping application that downloaded reviews for 1,000 companies.

In April 2020, we parsed and downloaded 2,587,686 reviews posted between 2011 and 2020. Along with the review, we collected other variables like overall rating, review title, employee position title, employee status(current or former, location, date of review, pros, cons, and the number of people that found review is useful or not.)

Since targeted technologies roles, a job title list[16] was used to filter the reviews. In order to filter for the job roles, a character-based tri-gram model was constructed with TF-IDF and measured to our job list using cosine similarity[19][17].

[14] https://business.linkedin.com/marketing-solutions/insight-tag
[15] https://blog.lempod.com/linkedin-industries-list/
[16] https://www.thebalancecareers.com/list-of-information-technology-it-job-titles-2061498
[17] https://towardsdatascience.com/fuzzy-matching-at-scale-84f2bfd0c536

Table 1 shows 3716 unique technologies roles distributed between industries and job titles with the highest number of reviews.

Table 1. Count of reviews and unique job titles for each industry

Industry	Reviews	Unique Job Titles
Accounting	369	194
Automotive	1090	338
Computer and Software	2624	636
Construction	228	121
Education Management	582	224
Financial	5151	1217
Higher Education	822	316
Hospital Care	2485	726
IT Services	29604	3012
Retail	2732	726

After data cleaning and finalizing technology job titles, 45251 reviews remained for topic modeling. The number of reviews scraped for each industry varies. This variability is likely due to the difference in the number of technology roles hired by each industry worldwide. IT services companies have almost 65% of total reviews ,and construction companies have only 0.5% of total reviews.

Employee reviews for these ten industries had a significant jump after 2012[2], doubling its reviews. This rise is expected because using social media has become common for job searching and that technology roles have started to be more in demand. As only 0.1 percent of reviews were posted in 2011, we noticed no reviews for 'higher-ed,' 'computer-software,' 'financial' and the 'automotive' industries (Fig. 10[)

Table 2. Percentage of reviews by year

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Percent of Reviews	0.1	3.5	7.4	9.3	9.8	11.1	25	15.5	13.8	4.5

Besides the number of reviews, each industry's rating average became almost constant from 2011 to 2020, and there are no visual differences in average rating across industries (Fig. [10]).

Except for construction, education management, and retail, the percentage of the reviews across the industries seems to be balanced between current and former employees (Tab. [3]). Word counts are shortest for IT on average ,and construction has the highest word count.

Table 3. Review Statistics by each industry

Industry	Average rating	Average review word count	Current employees' reviews (%)	Former employees' reviews (%)
IT services	3.8	38.2	49	51
Accounting	4	43.1	45	55
Automotive	3.8	42.4	43	57
Computer and software	3.8	44.2	43	57
Construction	3.8	56.8	37	64
Education management	4	47.5	37	63
Financial	4	42	42	58
Higher education	4	46.6	49	51
Hospital care	3.8	44	45	55
retail	3.7	47.7	36	64

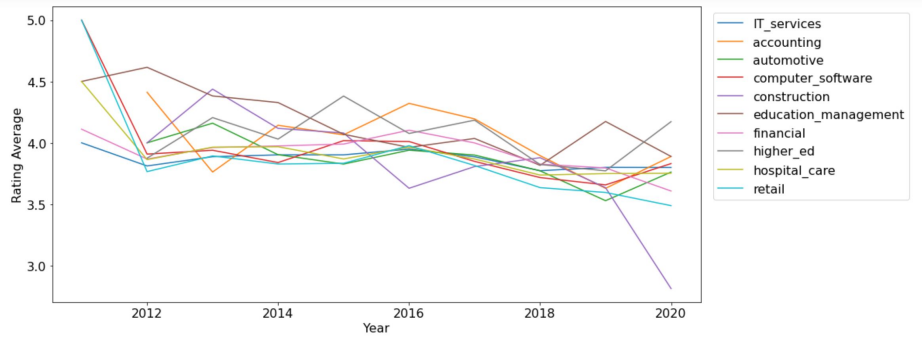

Fig. 10. Rating Average For Industries 2011-2020

5 Opinion Mining (Machine Learning Topic Modeling)

5.1 What is Opinion Mining

Opinion mining is used to extract people's opinions towards a topic or entity[10]. In our research, we wanted to obtain opinions from a niche job family to determine if there are new topics discussed by IT Professionals. The growing amount

of unstructured data, or reviews, makes it hard to mine data without advance topic modeling techniques.

Topic modeling is to find a similar group of words that create a single topic from a document and compare them to another document. As researchers, we can do this manually by inferring keywords like advancement and opportunities as career growth (Fig [11]). When applying this technique to thousands of reviews, manual annotation can be very resource intensive. Machine Learning techniques like Latent Dirichlet allocation (LDA) and Attention Based Aspect Extraction (ABAE) provides us tools to quickly divide our reviews into several defined topics based on all our reviews and provide informative insight and keywords.

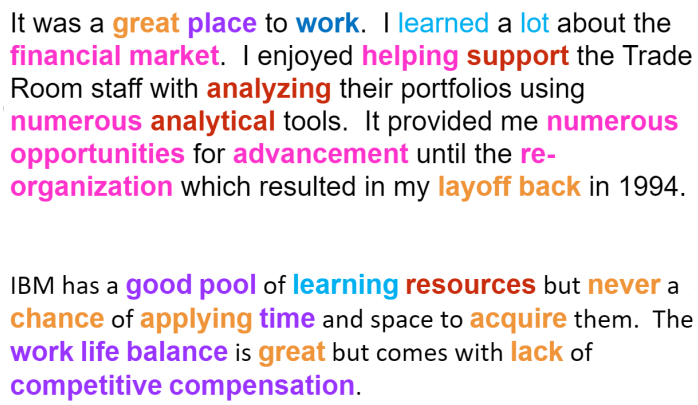

Fig. 11. Reviews from Indeed with highlighted keywords. The highlight colors are based on our final LDA model found in Section 7.

Two models were used to determine new topics discussed in company reviews, macro review levels, and micro sentence levels. The Macro-level analysis examined the entire review using Latent Dirichlet allocation (LDA). Attention Based Aspect Extraction analyzed individual sentences, micro-level analysis. Both approaches allowed us to identify each review's dominant topic and explore it in more detail at a subtopic level.

5.2 Latent Dirichlet allocation

Latent Dirichlet allocation was reintroduced in 2003 in the context of machine learning and presented as a visual tool for topic modeling. LDA is a probabilistic learning model where we can predefine a set number of topics. Topics are learned by creating a mixture of words and assumes that documents are a mixture of topics[3]. The model uses a generative statistical model that will sample the words from each topic and document. When the review is short, LDA tends to

experience problems from data sparsity[8]. We looked at a second model that performs better on a smaller corpus[7].

5.3 Attention Based Aspect Extraction

Neural Network has been extremely popular within Natural Language Processing (NLP), since the development of Word2Vec[13]. Word2Vec was trained on a shallow neural network that predicts the next word using a combination of skip-gram and a continuous bag of word design. Once trained, each word will be assigned a vector of a numeric value based on that corpus's common context. Researchers can use these vectors in text classification and unsupervised models like Attention-based Aspect Extraction (ABAE). Attention model is another technique used in NLP, that can be used for sentencing summarization[14]. ABAE, introduced in 2016, uses a neural network that includes a pre-trained Word2Vec embedding layer. The model uses an attention mechanism to de-emphasizes words not related to the aspect topic, thus creating topic words[7].

Both models are unsupervised, but requires a predefined number of topics. In the next section, we discuss how to fine-tune our models to determine the optimal number of topics.

6 Modeling

6.1 Data Preparation

Data preparation was performed in order to reduce the number of words within our vocabulary. Our initial corpus of 45k reviews contained 82k unique words. We identified 1,000 common misspelled words that were manually replaced. Lemmatization was used from the NLTK library[18] to replace multiple forms of a word with the base form, or lemma, of that word. The last technique we performed was removing high-frequency words or stop words and company names. Our final vocabulary contained 27k, unique words.

The vocabulary used for LDA was a combination of uni-gram and bi-gram. We excluded low and high-frequency words. For a word to be considered low or high-frequency, they must appear in more than five documents and less than 90% of the documents. ABAE used a Word2Vec embedding layer that was trained on all 2.5 million reviews. The parameters we selected to train our embedding layer were a vector size of 200, a window size of 10, and a minimum word count of five.

6.2 Hyper-parameters Tuning

Identifying the optimal number of topics is done through a series of parameter tuning and comparing the outcome to a topic coherence score[5].

[18] `https://www.nltk.org`

Table 4. Coherence score for top 5 LDA model

Model	alpha	beta	k	Coherence
1	asymmetric	0.91	13	-4.009
2	asymmetric	0.91	14	-4.135
3	symmetric	0.91	8	-4.240
4	0.61	symmetric	8	-4.274
5	symmetric	0.61	8	-4.364

The LDA parameters used to tune our model were document-topic density (α), topic-word density (β), and a number of topics (k). Gensim's[19] implementation of LDA was used for modeling and, the UMass Topic Coherence score[15] was used to evaluate the models (Tab [4]). Topic coherence allows us to measure relevant words within a topic.

ABAE tuning included negative samples (sentences), regularization term (λ), and a number of topics (k). The model was constructed using a modified version[20] of ABAE using TensorFlow 2.0. Each model was compared using a different implementation of UMass[12] (Tab [5]).

Table 5. Coherence score for top 5 ABAE model

Model	sentences	lambda	k	Coherence
1	20	0.10	5	-17.131
2	20	0.01	6	-17.372
3	40	0.25	5	-17.539
4	50	0.75	10	-17.601
5	40	0.25	8	-17.945

To learn which models were meaningful, we used human judgment to evaluate the top words[4]. LDA models containing thirteen and fourteen topics had a very low percentage of reviews, and in some cases, top words were nonsensical to IT professionals' interests, like hour, hike, and numbers. LDA model three was selected for its balanced distribution of words and low overlap of top words across topics. One of the topics contained only 2% of reviews and was merged with another topic with similar reviews. For our sentence analysis using ABAE, we found five topics to be the optimal model.

[19] `https://radimrehurek.com/gensim/index.html`
[20] `https://github.com/madrugado/Attention-Based-Aspect-Extraction`

7 Results

Thirteen topics were initially identified but were reduced to eleven useful topics. *Technical Skills* appeared in both of our models, and we merged two of the macro topics during our assessment of top words. The assignment of topic names was performed through our own opinion and an online survey. IT and non-IT professionals participated in a questionnaire of what they believed each group of words characterized.

Table 6. ABAE top words for each topic

Topic	% of Reviews	Top Words
Management	30.62%	management, work, manager, leadership, managing, staff, leader, cooperate, director, member, departmental, communicate, collaboration, ordinate, behavior
Technical Skills	19.94%	learn, platform, technique, knowledge, expose, technology, skill, fundamental, analytic, domain, networking, development, exploration, organization, enhance
Human Resource & Benefits	17.59%	salary, employee, sadly, bonus, pay, apparent, compensate, obviously, benefit, ethical, paid, deserve, wage, raise, actually
Work Life	16.75%	work, okay, ok, love, definitely, decent, probably, great, stay, summer, part-time, perfect, retire, live, start
Support	15.10%	ticket, report, log, verify, file, batch, manually, check, document, reject, monitor, confirm, release, final, request

Our final results can be found in tables [6] and [7] .

Indeed used five areas for employees and employers to research when reading reviews. By being niche, we were able to gain additional insight into company reviews. The topic models found three new topics and eight topics related to Indeed's list.

New topics found are *Learning & Development*, *Technical Skills*, and *Support*. Two of the new topics derived by our macro analysis, *Learning & Development* and *Technical Skills*, shows IT Professionals want to improve their skills indicated by a strong connection of the subtopics; *Technical Skills* and *Support* (Fig [12]).

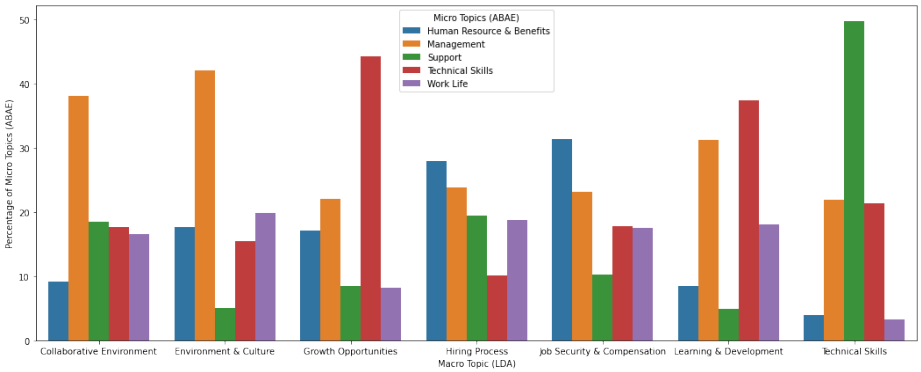

Fig. 12. ABAE/LDA Topics relationship

Employees in technology industries discussed education twice as much as employees in non-traditional technology industries. The large gap between IT and other industries and the associations between training and technical skills show IT firms promote learning new skills. There is also a relationship of having management involved in reviews related to *Learning & Development*.

Management, an Indeed topic, is discussed as a subtopic related to our more dominant topic from the macro analysis. Our micro analysis showed discussions about Management are specific to two topics *Collaborative Environment* and *Environment & Culture*. We found words related to collaboration to be around learn, help, and co-workers (Tab [7]), which Management can assist in improving.

Table 7. LDA top words for each topic

Topic	% of Reviews	Top Words
Hiring Process	25.18%	work, management, people, get, employee, great, not, job, year, manager, time, make, pay, like, hire
Environment & Culture	23.75%	work, good, place, balance, life, culture, environment, great, good work, management, work life, place work, life balance, employee, work culture
Learning & Development	19.83%	learn, work, good, lot, technology, place, experience, thing, get, great, opportunity, skill, learn lot, project, environment
Collaborative Environment	12.10%	work, part, enjoy, job, hard, part job, learn, help, time, management, hard part, always, typical, co-worker, enjoy part
Technical Skills	7.44%	system, development, software, support, application, work, project, use, test, client, customer, business, issue, network, process
Growth Opportunities	6.06%	train, provide, opportunity, employee, growth, skill, organization, technology , service, grow, world, professional, great, development, excellent
Job Security & Compensation	5.65%	project, good, salary, depend , less, hike, job, get, onsite , security, job security, growth , opportunity, fresher

Reviews showed that Management could also directly impact the culture of a team.

The topic *Hiring Process* was initially identified as *management*; however, through our survey process, we found it more of a holistic view of an employee's onboarding and hiring process. The relationship to its subtopics within those reviews shares a similar story that includes *Human Resources* & *Benefits* topics being discussed in 30% of the reviews related to onboarding and Management about 25% of the time(Fig. [12]). The absence of a dominant subtopic leads us to believe the primary topic is a holistic view of starting a career at a company. Non-traditional technology companies' top topic is *Hiring Process* (Tab [8]), and 30% of its reviewers rated the company 1 star. The low rating makes us believe that if a company improves its hiring process, they may increase their chances of recruiting top IT talent.

Another example of our focused topics compared to Indeed is *Growth Opportunities* alignment with *Job Security* & *Advancement*. *Growth Opportunities* having 45% of sentences related to subtopic *Technical Skills* shows a strong association between IT professionals learning new skills and advancing their career.

Our macro topic *Job Security & Compensation*, combines two Indeed topics, advancement, and pay, which were discussed the least.

Table 8. Macro topics between IT and Non-IT Industries

Topic	IT	Non-IT
Hiring Process	20.5%	36.4%
Technical Skills	6.4%	9.9%
Environment & Culture	25.8%	19.5%
Collaborative Environment	11.1%	14.5%
Job Security & Compensation	7.4%	1.4%
Learning & Development	23.1%	12.0%
Growth Opportunities	5.7%	6.3%

Employees working for Technology companies discussed pay seven times more than employees working in Non-Technology companies (Tab [8]).

Subtopic *Work Life* is a direct relationship to Indeed's topic. The sentences related to this topic are around a work schedule, and the sentiment reviewers had towards work life. *Work Life* never dominated any of the seven macro topics, indicating IT professionals do not focus on time and schedules.

8 Ethical Consideration

When it comes to ethics in research, we must first define ethics in the broader scope. Webster states, "ethics is a moral principle that governs a person's behavior or the conducting of an activity."Because of these "moral principles that govern,"we feel that there will always be ethical concerns in data collecting and data science research in general. Reviewing any bias' the data may have needs to be addressed. There are also ethical concerns that should be addressed when data is used in research that show specific results or could guide a user in making decisions.

Before beginning, a review at Indeed users must agree to their Review Guidelines. We assumed that users who filled out the original reviews and thus created the data we collected follow Indeed's guidelines. Due to the amount of data collected, we cannot verify if every user adhered to all of the Indeed Review Guidelines. We feel comfortable moving forward with an acceptance that enough reviewers have followed the guidelines close enough to not bias our overall findings.

Since data collected here could guide users in making decisions, we have been intentional in allowing all results collected through our research to be presented as is and has not been altered. All Graphs and charts presented throughout this report were prepared from that unaltered data.

9 Conclusion

Targeting a specific audience like IT Professionals allowed us to discover unique topics related to Technology. The two-level approach of topic modeling on reviews and sentences created insight into how topics relate to each other. We feel the information provided in this paper is a good starting point for non-traditional technology companies to refocus their strategy to attract IT talent. As an example, career advancement was driven by employees wanting to improve their technical skills. To attract new talent, companies should shift their learning to offer a more technical curriculum.

Based on the topic modeling research from company reviews, we built a base profile of how an IT professional perceives a company. Continuous education is highly sought after by IT talent. Education is one of the most discussed topics within the IT industry that wants to learn additional technical skills. Professionals expect management to be involved in a collaborative environment that promotes positive culture. Work life balance by IT professionals is emphasized over time and schedule. Finally, new associates look for a leadership supported on-boarding program. Companies that can foster these environments may improve their chance of attracting top IT talent.

Preliminary analysis of star ratings showed that the Hiring Process had the lowest rating, but no other sentiment analysis was performed. As research starts to glean topics unique to a job family, negative and positive sentiment should be considered. Sentiment analysis could help learn why employees may leave or continue their employment at a company.

Changes made, based on our contribution, could be evaluated through the metric time to fill. The next steps would be to work with HR and recruiters on promoting these topics with a longitudinal study. A study of this magnitude can determine if there is a significant change to the overall IT staffing of a non-traditional technology company.

References

1. Aruna, M., Anitha, J.: Employee retention enablers: Generation y employees. SCMS Journal of Indian Management **12**(3), 94 (Jul 1, 2015), https://search.proquest.com/docview/1721916370
2. Baird, C.H.: Myths about millennials in the workplace (Jan 1, 2018), https://search.proquest.com/docview/2002179329
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (January 3, 2003), http://www.jmlr.org/
4. Boyd-Graber, J.C.J.: Reading tea leaves: How humans interpret topic models
5. Clint P. George, H.D.: Principled selection of hyperparameters in the latent dirichlet allocation model
6. Guiso, L., Sapienza, P., Zingales, L.: The value of corporate culture. Journal of Financial Economics **117**(1), 60–76 (Jul 2015), http://dx.doi.org/10.1016/j.jfineco.2014.05.010

7. He, R., Dahlmeier, D., Lee, W.S., Ng, H.T.: An unsupervised neural attention model for aspect extraction. Tech. rep., SAP Innovation Center Singapore (Oct 11, 2018), https://www.comp.nus.edu.sg/~leews/publications/acl17.pdf
8. Hongqi, Q.X.H.Y.C.Z.L.X.T.J.H.T.W.: Burst-lda a new topic model for detecting bursty topics from stream text. **31**(6), 565–575 (2014), http://lib.cqvip.com/qk/85266X/201406/663771399.html
9. Jung, Y., Suh, Y.: Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. Decision Support Systems **123**, 113074 (Aug 2019), http://dx.doi.org/10.1016/j.dss.2019.113074
10. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1), 1–167 (May 23, 2012)
11. Luo, N., Zhou, Y., Shon, J.J.: Employee satisfaction and corporate performance: Mining employee reviews on glassdoor.com. In: Thirty Seventh International Conference on Information Systems, Dublin. pp. 149–167. John Wiley & Sons, Ltd, Chichester, UK (2016), https://pdfs.semanticscholar.org/b784/71ca2ac990e3ab2d70283e42e2bb5d3a8f7a.pdf
12. McCallum, E.T.M.L.A.: Optimizing semantic coherence in topic models
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (Oct 16, 2013), https://arxiv.org/pdf/1310.4546.pdf
14. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. Association for Computational Linguistics (ACL) (2015), https://search.datacite.org/works/10.18653/v1/d15-1044
15. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. pp. 399–408. WSDM '15, ACM (Feb 2, 2015), http://dl.acm.org/citation.cfm?id=2685324
16. Schramm, J., Mulvey, T.: New talent landscape: Recruiting difficulty and skills shortages. Tech. rep., The Society for Human Resource Management (June 2016), https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/SHRM%20New%20Talent%20Landscape%20Recruiting%20Difficulty%20Skills.pdf
17. Stamolampros, P., Korfiatis, N., Chalvatzis, K., Buhalis, D.: Job satisfaction and employee turnover determinants in high contact services: Insights from employees'online reviews. Tourism Management (12 2019)
18. Stamolampros, P., Korfiatis, N., Kourouthanassis, P., Symitsi, E.: Flying to quality: Cultural influences on online reviews. Journal of Travel Research **58**(3), 496–511 (Mar 2019), https://journals.sagepub.com/doi/full/10.1177/0047287518764345
19. Tata, S., Patel, J.: Estimating the selectivity of tf-idf based cosine similarity predicates. ACM SIGMOD Record **36**(2), 7–12 (Jun 1, 2007), http://dl.acm.org/citation.cfm?id=1328855
20. Walford-Wright, G., Scott-Jackson, W.: Talent rising; people analytics and technology driving talent acquisition strategy. Strategic HR review **17**(5), 226–233 (2018), https://search.datacite.org/works/10.1108/shr-08-2018-0071