European Proceedings of
Social and Behavioural Sciences
EpSBS

www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2020.03.03.69

ICMR 2019
8th International Conference on Multidisciplinary Research

TEXT ANALYTICS ON COMPANY REVIEWS IN JOBSTREET

Chiew Sook Chin (a), Gan Keng Hoon (b)*, Nur Hana Samsudin (c)
*Corresponding author

(a) School of Computer Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia, sookchin.chiew@student.usm.my
(b) School of Computer Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia, khgan@usm.my
(c) School of Computer Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia, nurhana.samsudin@usm.my

Abstract

During a job searching process, the condition of a company is always one of the considering factors to assist job seekers in their job selection process. Basically, job seekers are interested to know about the company environments, benefits as well as problems. In the old days, this information can only be obtained from company's employees if by any chance the job seekers know anyone in the company. Nowadays, the job seekers can also get to know the company environment through the reviews posted by company's employees. However, reading long reviews may be time consuming. Hence in this work, a text analytics pipeline (including cleaning, lemmatization and mining) is proposed to develop quick visuals solutions for job seekers to gain a quick insight about companies as well as making a quick comparison between companies if needed. Specifically, this study focuses on review contents of "the good things" and "the challenges" which represented positive and negative reviews about the company respectively. For evaluation, job reviews from three multinational companies were used. Feedbacks from the respective company's employees were obtained to verify whether the presented outcomes from the visuals are accurate and informative in representing positive (good things) and negative (challenges) about their companies.

Keywords: Text mining, n-gram, job reviews, information extraction, decision making.

https://doi.org/10.15405/epsbs.2020.03.03.69
Corresponding Author: Gan Keng Hoon
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Job searching is a common process experienced by most people regardless of fresh graduates or experienced workers. From job searching to successful recruitment could be a long and complicated process. With the advancement of internet, this process has become simplified and transparent where job seekers can perform their own job searching as well as evaluation based on the information contributed by public through the internet. However, it can be time consuming to read all the information.

Since the condition of a company is always an interest evaluation factor by job seekers during job searching, in this work, a text analytics technique, i.e., term frequency was used to develop a platform for job seekers to gain quick insight about their shortlisted companies. The data source used in this work is the company reviews posted by employees in JobStreet. Three multinational companies were selected for the study and evaluation of the reliability of the proposed technique. The outputs of this work are word clouds that represent the pros and cons of each company, which can provide quick insights about each company as well as offering comparison between companies.

2. Problem Statement

In general, there are few factors considered by job seekers during job searching process. Company related information is usually one of the most crucial factors to be considered. Today, performing company background check through internet is a common practice by many job seekers to gain better understanding about a company such as the company nature, business model etc (Stone, Neely, & Lengnick-Hall, 2018). During company background check, information about the company environment, company culture, employee's benefits as well as the comments from the existing employees are important information that the job seekers are looking for.

Traditionally, this type of internal information could be obtained through family, friend or friend's friend who have experience working with that company. However, the information gathered through this traditional way is limited. Nowadays, some job searching platforms such as JobStreet or Indeed allow employees to provide their personal comments about the company in the online platforms. These practices offer new opportunity for the job seekers to understand about a company from a whole new and genuine internal perspective rather than just solely depending on the information available on the company websites. (Schneider, Wickert, & Marti, 2017). However, it could be time consuming to review all the comments to identify the pros and cons about a company. In addition, a job searching process will usually involve more than one company. Therefore, an effective mechanism is needed assist in the review of all the employees' comments.

3. Research Questions

In this work, the goal is to propose a text analytics solution for job seekers that improve the efficiency of job searching process. The research questions to guide the work as follows:

a) How text analytics can be applied to improve job searching process?
b) What kind of texts are suitable for text analytics in job searching process?
c) What insights can be obtained from simple text analytics component like Word Cloud?

d) Can simple text analytics help to improve the targeted issue of job search process?

4. Purpose of the Study

The purpose of this work is to propose a text analytics pipeline (including cleaning, lemmatization and mining) that allow effective usage of the available information. This pipeline will produce visual solutions from information for job seekers to gain a quick insight about companies as well as making a quick comparison between companies.

5. Research Methods

In this section, the pipeline of text analytics (on the company reviews posted in JobStreet) is proposed to offer a quick insight about the shortlisted companies. The tools using in this work were Jupyter Notebook and Python in Intel IA Dev Cloud. Online NGram Analyzer was also used for certain checking or testing. The overall flow of text analytics is as below (Refer Figure 01).

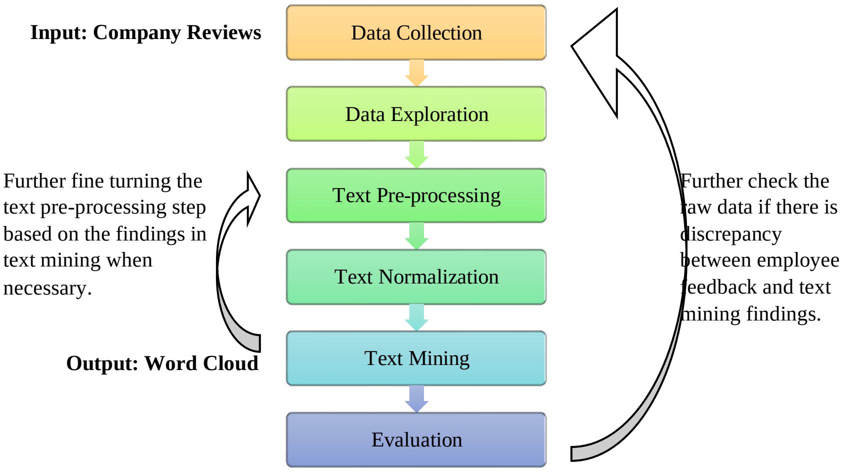

Input: Company Reviews
Data Collection
Data Exploration
Text Pre-processing
Text Normalization
Text Mining
Evaluation
Output: Word Cloud
Further fine turning the text pre-processing step based on the findings in text mining when necessary.
Further check the raw data if there is discrepancy between employee feedback and text mining findings.

Figure 01. Flow of text analytics used in company review analysis

5.1. Data Collection

The data using in this work are company reviews posted by companies' employees in JobStreet website. At the initial stage of work, three well-known companies in Penang were selected, namely Company A, Company B and Company C. A total of 132 reviews posted from 01 Jan 2017 to 10 Dec 2018 were collected from JobStreet website (JobStreet.com Malaysia, 2018) and the review distribution of the companies is shown in Figure 02.

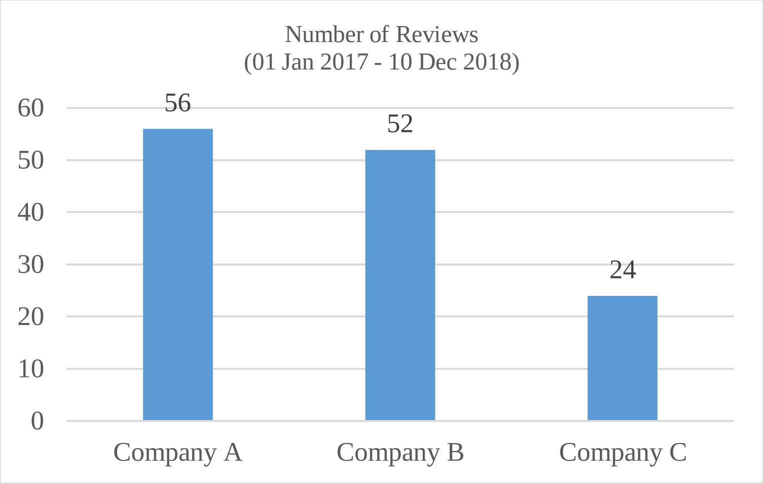
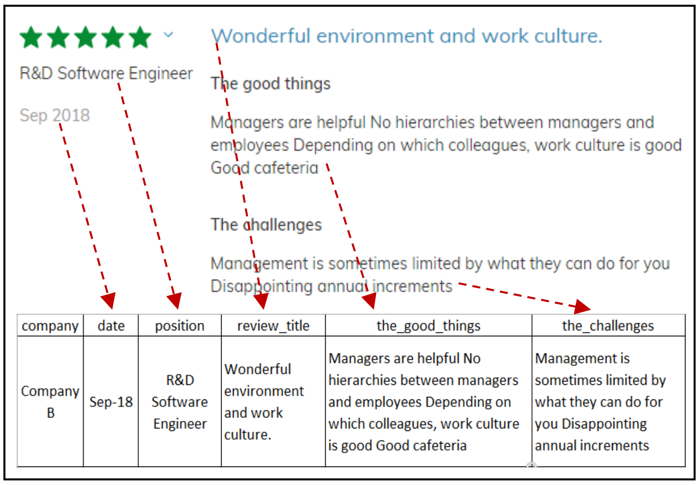


Figure 02. The distribution of collected reviews by companies

The data size might be insufficient to provide a comprehensive information about the company but it should be sufficient to provide basic significant insight. The company reviews collected from JobStreet website were transformed into Excel table for further processing. The data collection and transformation process are summarised in Figure 03. As shown in Figure 03, there are two parts of reviews, i.e., "the good things" and "the challenges" which represented the positive and negative reviews about the company respectively. Thus, separate analysis was required for these two parts of reviews.



★★★★★

R&D Software Engineer

Sep 2018

Wonderful environment and work culture.

The good things

Managers are helpful No hierarchies between managers and employees Depending on which colleagues, work culture is good Good cafeteria

The challenges

Management is sometimes limited by what they can do for you Disappointing annual increments

company	date	position	review_title	the_good_things	the_challenges
Company B	Sep-18	R&D Software Engineer	Wonderful environment and work culture.	Managers are helpful No hierarchies between managers and employees Depending on which colleagues, work culture is good Good cafeteria	Management is sometimes limited by what they can do for you Disappointing annual increments

Figure 03. Collect company reviews and transform to Excel table

5.2. Data exploration

After the data collection, exploration and inspection were performed to ensure the validity of the data collected. There were 132 rows and 6 columns in the collected data. The 132 rows are the reviews for the three companies. The descriptions for the six columns are shown below (see Table 01). Data exploration on data type for each column was as performed to make sure no inappropriate format is used, as well as no missing value. All collected data are stored in Excel table. After that, the dataset is ready for pre-processing stage.

Table 01. The descriptions of the companies

Column name	**Description**
Company	Name of company
Date	Month and year of review posting
Position	Position of employee
Review_title	Title of review
The_good_things	Positive review about the company
The_challenges	Negative review about the company

5.3. Text Pre-processing

In this stage, data were categorized as "the_good_things" and "the_challenges" which represented the positive and negative insights of company, respectively. Since the data is unstructured data, therefore data cleaning is required. Based on Loper and Bird (2002), there are six steps taken in the text pre-processing process: first, tokenize the text data by using unigram tokenization; second, remove the punctuations in text data; third, remove extra whitespace between each word for standardization; fourth, convert the text to lowercase for standardization purpose. However, exception for certain words such as "US" and "IT" were applied to avoid the deviation of the word meanings. The list of exception words was designed in scalable format in Python script where new exception word can be added easily in future; fifth, remove numeric text which is not meaningful in representing the insight about company; and finally remove stop words that have no significant meaning such as "the" and "of". The stop word list is a predefined list in Python Natural Language Toolkit (NLTK) library. In terms of tokenization, this work uses the unigram tokenization as initial evaluation (to check the feasibility to apply text analytics in job searching process). A simple testing was also performed in Online NGram Analyzer using bi-gram or more tokenization to check if there are any significant outputs. The findings will be further discussed in Text Mining section.

5.4. Text Normalization

After the text pre-processing step, the cleaned text data require additional step for standardization. The wording or format standardization step is called text normalization. In general, there are two types of normalization, which are stemming and lemmatization. Stemming is a process to convert the word to its stem which is a good method to standardize different wording with same stem. However, stemming was not used because it would produce invalid wording which could bringing difficulty to understand the outputs of analytics (word clouds). Lemmatization is another method to standardize the text. The output of lemmatization is valid word which overcome the issue in stemming. It might not as good as stemming in standardizing the word to stem word, but it is good enough in basic wording standardization such as plural convert to singular. Thus, lemmatization is applied to normalize the text data in this work.

5.5. Text Mining

In order to identify the significant insights from the company reviews, basic text analytic technique which is term frequency or word count has been applied. An assumption is that the frequent term in the reviews would be more likely representing the insight about the company. There are two types of graphics used to represent the frequent words: word cloud and histogram. The word cloud provides quick glance insight about the company for job seekers. Meanwhile the histogram visualize the quantitative information about word frequency. This graphical output was only be used as a reference in the development stage, not to deliver as final output for job seekers.

As the objective of work is to provide a quick insight about positive and negative of each company, thus, there would be different word clouds to represent positive and negative reviews. Since there are three companies were selected for in this study, a total of six word clouds will be generated to represent the positive and negative comments about the three companies.

In the analysis of preliminary graphics developed from normalized text data, words with less significant meanings were found. These less meaningful words such as “also” and “may” were categorized as stop words, which was then removed during the text pre-processing stage. Besides, it was also found that the preliminary word cloud was rather crowded as showed in Figure 04, which caused distraction to identify the meaningful keyword.

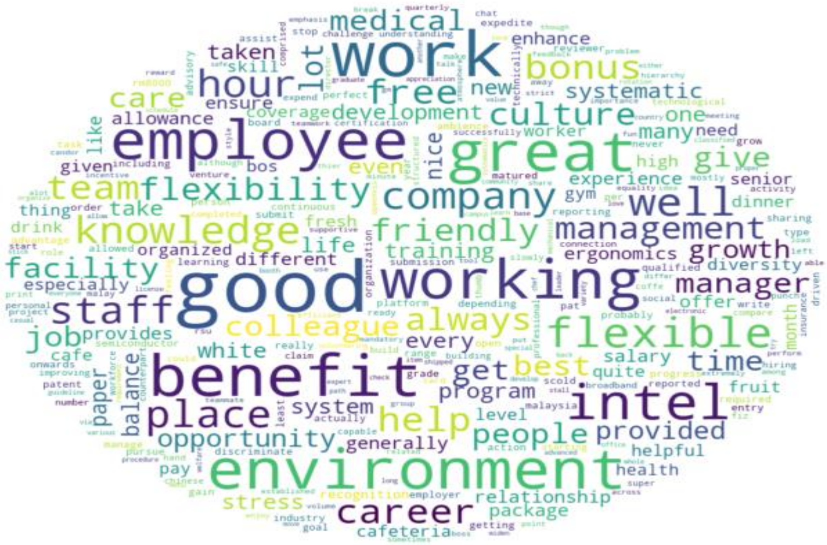


Figure 04. Preliminary word cloud without applying term reduction

Thus, term reduction was applied to reduce the term with frequency of 2 or below as it is least likely to represent the company insight. However, only 24 reviews were available for Company C which is half the size of the reviews available for Company A and Company B. After removed the term with frequency ≤ 2 for Company C, it was only left less than 20 terms. Thus, the term reduction for frequency ≤ 1 was applied for the reviews of Company C. In overall, 83.7% of terms were reduced as shown in Table 02.

Table 02. Term Reduction Comparison

Company	Type of Review	Number of Term		% of Reduction
		Before	*After	
Company A	the_good_things	312	58	81.4%
	the_challenges	375	47	87.5%
Company B	the_good_things	241	43	82.2%
	the_challenges	293	31	89.4%
Company C	the_good_things	168	39	76.8%
	the_challenges	141	32	77.3%

https://doi.org/10.15405/epsbs.2020.03.03.69
Corresponding Author: Gan Keng Hoon
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Company	Type of Review	Number of Term		% of Reduction
		Before	*After	
	Total	1531	251	83.7%

*Note: Company A & Company B apply reduction for term frequency ≤ 2; Company C apply reduction for term frequency ≤1.

By removing the identified less meaningful word and apply term reduction, the new word clouds have been developed and illustrated in Table 03.

Table 03. Word clouds for company reviews in JobStreet

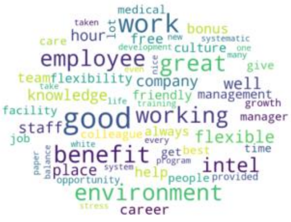
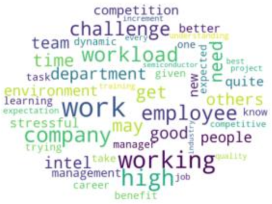
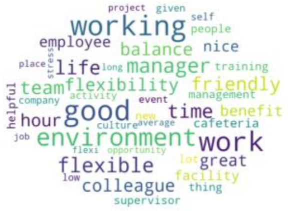
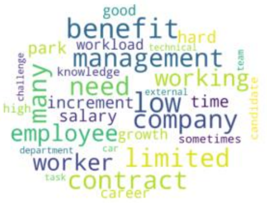
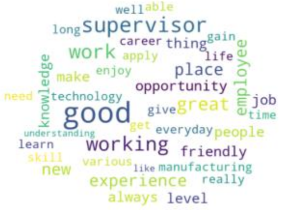
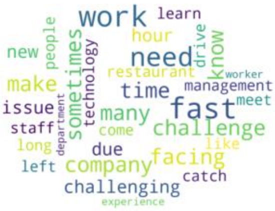
Company	The Good Things	The Challenges
CompanyA		
Company B		
Company C		

Table 04 below shows some of the positive (good things) and negative (challenges) insights about the company obtained from the six Word Clouds.

Table 04. The positive and negative insights obtained from word cloud

Company	The Good Things	The Challenges
Company A	▪ Great working environment ▪ Benefit ▪ Flexible time ▪ Bonus ▪ Well company management ▪ Good facility	▪ Competitive ▪ Workload ▪ Stressful ▪ High expectation
Company B	▪ Good working environment ▪ Work like balance ▪ Flexible time ▪ Cafeteria	▪ Low salary ▪ Low increment ▪ Contract worker ▪ Company management

	▪ Friendly colleague	▪ Limited parking
Company C	▪ Good supervisor ▪ Great opportunity ▪ to learn new skill and experience ▪ Friendly employee	▪ Fast and challenging ▪ Company management ▪ Long drive to work ▪ Restaurant

As mentioned in Text Pre-processing section, a quick analysis is conducted using Online NGram Analyzer (Ngram Analyzer, 2018) by applying bi-gram or more tokenization to explore if there is any additional information. There are some frequent and significant ngram term found as showed in Table 05. The table shows that bi-gram or tri-gram tokenization can produce more specific and new insight such as career growth in Company A, helpful manager and the challenge of converting from contract worker to a permanent position in Company B.

Table 05. The frequent ngram term from online ngram analyzer

Company	The good things	The challenges
Company a	▪ Flexible working hour ▪ Work life balance ▪ Great working environment ▪ Career growth	▪ You are expected to ▪ Quite stressful
Company b	▪ Work life balance ▪ Flexible working hours ▪ Managers are helpful ▪ Environment is good	▪ Hard to convert to ▪ External contract worker ▪ To permanent ▪ Salary increment ▪ Career development
Company c	▪ Good experience	▪ Nil

However, Online NGram Analyzer is not a good platform to achieve the objective of this work due to the manual step needed to input the text, plus there is no text normalization feature (e.g. stemming and lemmatization). Online NGram Analyzer was just used as a quick check to assess whether different n-gram provide additional insights about the company in this work. Thus, expanding the unigram word cloud developed by Python to cover bi-gram or more tokenization is a potential future enhancement.

6. Findings

In this work, qualitative evaluation method is used to evaluate two components: reliability of the data collected from JobStreet, and usability of word cloud platform to represent the company’s insight. In the evaluation part, the feedbacks from some company’s employees have been collected to verify whether the word clouds in Table 03 are accurate and informative in representing positive (good things) and negative (challenges) about their companies. The summary of collected feedbacks is shown in Table 06.

Table 06. The feedbacks collected from employees

Company	Good Things			Challenges		
	√	X	Accuracy	√	X	Accuracy
Company A	3	0	100%	2	1	66.7%
Company B	3	0	100%	3	0	100%
Company C	2	0	100%	2	0	100%

√: *The information in word cloud is correctly representing the company*
X: *The information in word cloud is not good enough to represent the company*

Overall, the company's employees agreed that the word clouds correctly portray the goods and challenges of their companies. However, there are some minor specific feedbacks from employees which were different from the insights gained from word cloud. It was further check and discussed in Table 07 below:

Table 07. Specific feedbacks from employees and future check

Word Cloud	Feedback	Future Check
Company A – the challenges	Less accurate because the correct information is affected by the irrelevant word which in bigger size. In addition, the diversity issue caused by employees in different country, working hour and culture was not reflected in cloud.	▪ Irrelevant word could be solved by applying bi-gram or tri-gram tokenization in future. ▪ By checking on the raw data, diversity did not highlight in company review. More data require to represent comprehensive insight about the company in future.
Company B – the good things	The cafeteria is not considered as a good thing in Company B.	By checking on the raw data, the a few comments on cafeteria is good which the word cloud represented the reviews accordingly. However, it is a subjective judgment. It would be convincing if more data and employee feedbacks collected for development and evaluation.
Company B – the challenges	Limited parking is not an issue in Company B.	Future check on raw data indicated that limited parking is experienced by contract worker. Hence, the word cloud represented this issue correctly but not specifying to contract worker. It could be solved by applying bi-gram or tri-gram tokenization in future.

6.1. Discussion

The objective of work is to provide a platform for job seekers to gain quick insight about the company. In this work, word clouds used as platform to provide such insight to the job seekers. Overall, the word cloud (refer to Table 03) is able to produce accurate and informative insight as evaluated by respective employees. However, there are still potential gaps to be improved on the accuracy or to gain more insight by focusing on the factors and their potential future work (see Table 08).

Table 08. The factors for accuracy improvement and potential future works

Factor	Description	Future Works
Size of data	There was only 132 reviews collected from JobStreet website. Based on the findings, it is sufficient to reveal the general information about the companies such as good environment and stressful. However, more data was required for comprehensive or specify information.	▪ Collect company reviews from different sources such as Indeed.
Small sampling evaluation	There was only a small sample collected to evaluate the outputs which could not representing the overall organization.	▪ Collect more evaluation feedbacks from different department and level of position.
Unigram tokenization	The unigram tokens may not provide complete information. When glance on word cloud, guessing is needed in identifying meaningful phrase by combine 2 or more unigram token.	▪ As discussed in Text Mining section, expand the unigram to cover bi-gram or more tokenization can produce more and specify information in word cloud. ▪ Correlation between tokens would be a good method in identifying high correlated words for determine meaningful phrase. Correlation checking was not applied in this work due to technical limitation.

There are some challenges faced in this work such as size of data etc. Thus, future enhancement is needed to produce more accurate and informative outputs. As a preparation for future enhancement, the Python script was designed in a user friendly and scalable structure. The stop word list and lowercase exception list is expandable from time to time for newly detected word.

Other than job seekers, this solution (word clouds) is also applicable for company human resource (HR) personal for better understand the employee perspective and take necessary action as an employee retention approach. Besides, the HR personal can also understand the benefit offered by competitor companies in analysing the employee's needs and leaving reasons.

7. Conclusion

In conclusion, the proposed text analytics solution and its output (word clouds) has been evaluated by expert (company employee). In overall, the word clouds can serve as a platform to provide overall and quick insight about the company. The quick insight is a good reference and save time for job seekers in job searching process. It is also a good platform for HR personal to study and understand the employee's need and the challenges facing. In this work, the company reviews presented in word clouds are sufficient to provide quick glance on the main insight about the company. However, the improvements such as more data, expand the unigram tokenization are needed for more accurate and informative insights.

Acknowledgments

This research is supported by Fundamental Research Grant Scheme (FRGS), Ministry of Education Malaysia (MOE) under the project code, FRGS/1/2018/ICT02/USM/02/9 and title, Automated Big Data Annotation for Training Semi-Supervised Deep Learning Model in Sentiment Classification.

References

JobStreet.com Malaysia (2018). *Company Profiles and Reviews - JobStreet.com Malaysia.* Retrieved from: https://www.jobstreet.com.my/en/companies/xxxxxx/reviews

Loper, E., & Bird, S. (2002). Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. https://doi.org/10.3115/1118108.1118117

Ngram Analyzer (2018). Retrieved from: http://guidetodatamining.com/ngramAnalyzer/

Schneider, A., Wickert, C., & Marti, E. (2017). Reducing complexity by creating complexity: A systems theory perspective on how organizations respond to their environments. *Journal of Management Studies*, *54*(2), 182-208.

Stone, C. B., Neely, A. R., & Lengnick-Hall, M. L. (2018). Human resource management in the digital age: Big data, HR analytics and artificial intelligence. In *Management and technological challenges in the digital age* (pp. 13-42). CRC Press.