Text mining in career studies: Generating insights from unstructured textual data

Kobayashi, V.B.; Mol, S.T.; Vrolijk, J.; Kismihók, G.

DOI
10.4337/9781788976725.00015

Publication date
2021

Document Version
Final published version

Published in
Handbook of Research Methods in Careers

7. Text mining in career studies: generating insights from unstructured textual data[1]

Vladimer B. Kobayashi[2], Stefan T. Mol, Jarno Vrolijk, and Gábor Kismihók

Text data pertaining to peoples' careers have proliferated in the past few decades. Due to the digitization of job search, recruitment, and the development of HR systems, it is relatively easy to access and obtain large datasets containing information about jobs or other work-related information at the micro (individual), meso (institutional), and macro (regional, national and global) levels, or some combination thereof. Examples of text data that may be used to study careers include (auto)biographies, résumés, posts in professional social networking sites, online job boards, public surveys, interview transcripts, personal diary entries, and even academic publications. Of particular interest are job vacancies, as aside from education and job experience, they also contain information about individuals' roles, responsibilities, knowledge, skills, and abilities, which comes with the promise of adding specificity and context to the career domain, which has come to be dominated by reductionist and generalist approaches to operationalizing key constructs. Online forums and social media also provide data relevant to the study of careers since employees use these platforms to voice their ongoing opinions and sentiments about their past and present employers.

As a way to characterize big text data we can use the framework of the four "V"s of Big Data: Volume, Velocity, Variety, and Veracity (De Mauro, Greco, & Grimaldi, 2015). The sheer *Volume* of the available text data on careers is unprecedented, and far beyond the traditional qualitative and quantitative datasets in careers research. It is oftentimes not possible to store these data locally on a single computer (e.g. a desktop) and to use traditional analytical software and methods for their analysis. Furthermore, the rate at which data about work and careers is generated (*Velocity*) is also growing. One should simply think about the number of public status or CV updates on popular professional/social networking sites such as LinkedIn or Facebook, or the number of vacancy announcements posted to the Internet on a daily basis. Data also comes in many different guises (*Variety*), and are hardly ever produced with the primary aim of facilitating the conduct of research. Therefore, substantial effort must be invested to process different data types and forms in order to make the data suitable for analysis. A final challenge lies in the question of data and data-source integrity (*Veracity*), which also needs to be carefully considered when one wants to generate valid insights from textual data.

The abundance of "big" text data containing information about careers also offers new avenues for careers research and paves the way for the development of bespoke

text analysis methods and applications. Gone are the days when text analysis was limited to the mere counting of words, as contemporary sophisticated methods allow researchers to extract and organize textual content into topics, themes, and semantics, opening the door to theory generation and theory-testing (Kobayashi et al., 2018b). The current chapter was written to illuminate some of these possible avenues and to provide initial guidance to careers researchers who might be interested in working with "big" text data.

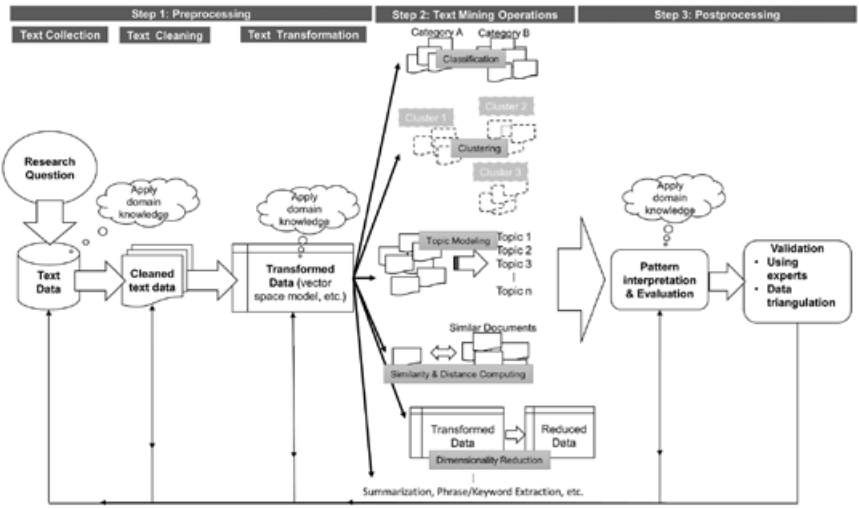
Most of the aforementioned types of text data are relatively easy to collect, as they can be simply scraped from the Internet. However, in some cases researchers may need to consult with website owners, and ask for direct access (for instance through an Application Programming Interface or API). It comes without saying that there are privacy and possibly copyright issues associated with the use of such data. Privacy and copyright considerations should be negotiated prior to data collection and analysis (van Wel & Royakkers, 2004).

In an effort to familiarize careers researchers with text mining practices, this chapter provides an overview of the text mining process, from data preprocessing, the application of text mining operations, validation, to post-hoc analysis of the resulting models. Various text mining operations, such as dimensionality reduction, text clustering, topic modeling, text classification, and neural word embeddings are discussed. Finally, to demonstrate the capability of novel text mining methods in the area of careers research, an example use case is provided.

TEXT MINING

The primary aim of text mining (TM) is to generate (exploratory) insights from and/or test hypotheses using free unstructured big text (Kao & Poteet, 2007). Hence, text mining is viewed as an objective-driven, systematic process that is generally performed through the following three steps: (1) text data collection and preprocessing, (2) application of text mining techniques, and (3) postprocessing (Zhang, Chen, & Liu, 2015). Figure 7.1 shows an overview of the different steps in the TM process. The organization of this chapter follows this figure.

Text preprocessing may be further subdivided into *text data cleaning* and *text data transformation* (e.g. converting unstructured text into intermediate forms which are used as input to the actual TM operations). TM operations refer to the application of algorithms with the goal of extracting hidden patterns and characteristics from text. Finally, postprocessing involves interpreting and validating knowledge obtained from TM operations. The focus of this chapter is on the different text mining operations. The readers are referred to other open source papers (Kobayashi et al., 2018b, 2018a) for a more comprehensive treatment of the other steps. Hereafter, all mentions of "data" refer to text data. Moreover, "document" and "text" are used interchangeably and corpus refers to the collection of text.



Source: Reprinted from Kobayashi et al. (2018b). Retrieved from https://doi.org/10.1177/1094428117722619. Licensed under a Creative Commons Attribution-NonCommercial 4.0 License.

Figure 7.1 Flowchart of the text mining process

Text Transformation: From Unstructured to Structured Data

Text mining techniques require that the input data be in a specific format. Before applying TM techniques, unstructured text data is first transformed so that TM techniques can be applied (Weiss, Indurkhya, & Zhang, 2015). Unlike common approaches to analyzing quantitative data, in which data cleaning, data transformation, and data analysis are sequential and separate stages, the specific TM operation determines the text transformation because oftentimes a TM operation accepts only a particular text data format. In fact, when subsequent results are unsatisfactory, one can try different combinations of representations (Lewis, 1992; Scott & Matwin, 1999), different transformation methods, or TM operations. Usually different combinations of data transformation and analytical techniques are tested and evaluated. In essence, text transformation is a representation strategy in which free text is converted into structured format (or formally mathematical objects). Most analytical techniques accept a matrix structure, where the columns are the variables (more commonly referred to as *features*) and the rows are the documents (for instance résumés, vacancies, or biographies). A straightforward approach is to construct this matrix by simply using the words (or *terms* as they are more commonly known) as variables. The resulting matrix is called a "document-by-term matrix" in which the entries of the matrix are raw frequencies of terms occurring in the documents. In many applications, this is an obvious choice since words are the basic linguistic units that

express meaning. Thus, in this transformation, each document is transformed into a "vector," the size of which is equal to the size of the vocabulary (i.e. the set of unique words in the corpus), with each element representing the number of times a particular term occurs in that document (Scott & Matwin, 1999). By using raw frequencies, commonly occurring words in documents are given more weight.

Term frequency, in itself, may not be useful if the task is to make groupings or categories of documents (Kobayashi et al., 2018a). Consider the word "study" in a corpus consisting of abstracts of scientific articles. If the objective is to categorize the articles into topics or research themes then this term is not particularly informative in this particular context since many abstracts will contain this word. A way to prevent the inclusion of terms that possess little discriminatory power is to assign weights to each term that reflect its specificity to particular documents in a corpus (Lan et al., 2009). The most commonly used weighting procedure is the Inverse Document Frequency" (*IDF*) (Salton & Buckley, 1988). It is computed using the following formula:

$$\mathrm{IDF}\left(\text{term i}\right) = \log \frac{N}{n}$$

where N is the corpus size and n is the number of documents containing term i.

A term with an *IDF* of zero is useless in the discrimination process because that term is present in every document (i.e. $N=n$). In fact, *IDF* can also be used as a criterion to filter out common terms, that is, terms that have a low *IDF* have little discriminatory power and hence can be disregarded. When the raw[3] term frequency (*TF*) and *IDF* are multiplied together it yields the popular *TF-IDF*, which in principle simultaneously accounts for both the word's frequency and specificity (Frakes & Baeza-Yates, 1992) with higher values being more desirable.

One disadvantage of representing texts as a document-by-term matrix, is that it ignores word order information. This can be especially problematic when the goal is to extract semantics in text (Harish, Guru, & Manjunath, 2010). However, it turns out that despite ignoring word order information, this representation seems to work well for many text classification applications, such as in email spam detection, document authorship identification, and topic classification of news articles (Song, Liu, & Yang, 2005; Zhang, Yoshida, & Tang, 2008). Another disadvantage of the document-by-term matrix representation is the resulting high dimensionality, that is, size of the vocabulary. One can use different data dimensionality reduction methods to reduce the number of variables (e.g. variable selection and variable projection techniques) or employ specific techniques suited for data with high dimensionality. These techniques will be discussed in the Text Mining Operations section.

There are other ways to represent text such as using the *n*-gram approach which uses *n* consecutive words or letters (including spaces) as features. Another is one-hot encoding which represents each word as a vector consisting of 1 in the position of the word and 0 in the other positions. The vector has the same size as the vocabulary. For example, suppose that the vocabulary only has these words {"the", "covid-19",

"pandemic", "killed", "thousands", "of", "people"} then the vector representation for "pandemic" is (0,0,1,0,0,0,0). This type of representation stands at the basis of neural word embedding (more on this later).

Once text is transformed, techniques such as classification or cluster analyses can be applied. Concatenating noncontiguous words can also tackle substantive questions about the text. For instance, in résumés of job applicants the proximity of the words "experience" and "year" together with a number are used to deduce applicants' length of work experience.

Text Mining Operations

In the following sections, those TM operations which we find most applicable to careers research are discussed. This is followed by a discussion of methods that may be used to assess the credibility and validity of TM outcomes. Most of the techniques covered here accept the document-by-term matrix as input, otherwise we shall explicitly mention the format of the input data.

Dimensionality reduction

Document-by-term matrices tend to have many variables and subsequent analyses may suffer from what is called the *curse of dimensionality* (Aggarwal & Zhai, 2012; Alpaydin, 2014). It is usually desirable to reduce the size of these matrices by applying dimensionality reduction techniques. Benefits of reducing dimensionality include more tractable analysis, greater interpretability of results (e.g. it is easier to interpret variable relationships when there are few of them), and more efficient representation. Compared to working with the initial document-by-term matrices, dimensionality reduction may also reveal latent dimensions (e.g. higher level concepts) (Yarkoni, 2010) and may result in improved performance (Bingham & Mannila, 2001).

Two general approaches are commonly used to reduce dimensionality. One is to construct new latent variables and the second is to eliminate seemingly irrelevant variables. New variables are modeled as a (non)linear combination of the original variables and may be interpreted as latent constructs. For example, the words "flexible", "willingness", and "abroad" may be merged to express the concept of a "willingness to travel" in a corpus of job vacancies. An added advantage of dimensionality reduction techniques is that they may be used to eliminate variable multicollinearity.

Singular Value Decomposition (SVD) is a classic tool which underlies techniques such as Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) and Principal Components Analysis (PCA) (Jolliffe, 2005). Reducing the number of dimensions is accomplished by retaining only the first few largest singular values. Usually, this implies choosing latent dimensions and recovering the right dimensionality of the data because at times, true dimensionality is obscured by random noise.

Latent Semantic Analysis (LSA) is commonly used when synonymy (i.e. different words that have the same meaning) and polysemy (i.e. one word used in different yet related senses) are present in the data. PCA is effective for data reduction as it preserves the variance in the data. Parallel analysis (Ford, MacCallum, & Tait, 1986;

Hayton, Allen, & Scarpello, 2004; Montanelli Jr & Humphreys, 1976) is the recommended strategy to choose how many dimensions to retain in PCA. A disadvantage of both LSA and PCA is that it may be difficult to interpret the derived dimensions. Another technique is Random Projection (RP) where data points are projected to a lower dimension while maintaining the distances among points (Vempala, 2005). Compared to PCA, RP is computationally less demanding and its results are comparable to those of PCA (Bingham & Mannila, 2001).

An alternative approach to reduce dimensionality is to eliminate variables by using variable selection methods (Guyon & Elisseeff, 2003). In contrast to projection methods, variable selection methods do not create new variables but rather select from the existing variables by eliminating those that are uninformative or redundant (e.g. words that occur in too many documents such as "the", "to", "and", "of"). Three types of methods are available: filters, wrappers, and embedded methods. Filters assign scores to variables and apply a cut-off score in order to select relevant variables. Popular filters are TF-IDF thresholding, Information Gain, and the Chi-squared statistic (Forman, 2003; Yang & Pedersen, 1997). Wrappers select the best subset of variables depending on the particular analytical method that is to be applied. Searching for the best subset of variables using embedded methods is accomplished by minimizing an objective function that simultaneously takes into account model performance and complexity. Model performance can be measured, for example, by prediction error (in the case of classification), and complexity is operationalized in terms of the number of variables in the model. In line with Ockham's razor, the preferred subset is the one that achieves the best balance between the number of variables (fewer is better) and prediction error (lower is better). In practice, the model prediction error is computed using a separate test set. That is, the model is developed on training data and validated on a separate sample of testing data.

Text Clustering

Many tasks in TM involve grouping documents such that documents belonging to the same group are similar and documents from different groups are dissimilar (Jain, Murty, & Flynn, 1999; Steinbach, Karypis, & Kumar, 2000). The process of grouping is called *clustering*. The main uses of text clustering are to organize documents to facilitate efficient search and retrieval and to impose an automatic categorization of documents. For example, text clustering has been used to create topical groupings in a collection of legal documents (Conrad et al., 2005) and automatic grouping of search query results (Osinski & Weiss, 2005). In many clustering procedures the researcher needs to define a measure of distance between texts. Commonly used measures that operate on vector representations are the Euclidean and Hamming distances.

Most clustering algorithms are categorized as either hierarchical or partitional (Steinbach et al., 2000). Hierarchical clustering algorithms are classified into either agglomerative or divisive. In agglomerative clustering, initially there are as many clusters as there are documents and then gradually clusters are merged until all

objects belong to a single cluster. Conversely, the divisive approach entails first assigning all documents to a single big cluster and recursively splitting clusters until each document is in its own cluster. The merging (or splitting) of clusters is oftentimes depicted using a tree or dendrogram.

For partitional clustering the user has to specify the number of clusters beforehand and clusters are formed by optimizing an objective function that is usually based on the distances of the objects to the centers of the clusters to which they have been assigned. The popular *k*-means algorithm is an example of partitional clustering (Derpanis, 2006). One key challenge in clustering is the determination of how many clusters to form. Since clustering is an exploratory and inductive technique, a common strategy is to try out different numbers of clusters (*k*) and use cluster evaluation metrics (e.g. Dunn index, Silhouette coefficient, or an external evaluation criterion) to decide upon the most suitable number of clusters (Jain et al., 1999).

Topic Modeling

Topic modeling can be applied to automatically extract topics from documents, where extracted topics represent latent constructs or themes. For example, in a corpus of exit interviews, one could set out to extract the various reasons that people have mentioned for quitting their current job. In machine learning and natural language processing, topic models are probabilistic models that are used to discover topics by examining the pattern of term frequencies (Blei, Ng, & Jordan, 2003). Its mathematical formulation has two premises: a topic is characterized by a distribution of terms and each document contains a mixture of different topics. The most likely topic of a document is therefore determined by its terms. For example, when an exit interview contains words such as "pay", "compensation", "salary", and "incentive", one of its candidate topics is "rewards or compensation".

Perhaps the most popular topic models are the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model and the Correlated Topic Model (CTM) (Blei & Lafferty, 2007). LDA and CTM both operate on the document-by-term matrix (Porteous et al., 2008). CTM will yield almost the same topics as LDA. The main difference between the two is that in LDA, topics are assumed to be uncorrelated (i.e. orthogonal), whereas in CTM topics can be correlated. In comparing LSA with LDA, the latter has been found to be particularly suitable for documents containing multiple topics (Lee et al., 2010). Supervised Topic Modeling is an extension of LDA (McAuliffe & Blei, 2008). The modeling assumptions remain the same as in LDA except that it is possible to incorporate an outcome variable. For example, one could use Supervised LDA to extract those career shocks from interview transcripts that are related to a quantitative measure of career sustainability.

Classification

Classification is the assignment of objects to predefined classes or categories, which unlike clustering are known a priori. Logistic regression is perhaps the best-known

classification method. The goal is to construct a model that can predict the category of a given document. Example applications of text classification are spam or ham classification of emails (Youn & McLeod, 2007), authorship identification (Houvardas & Stamatatos, 2006), thematic categorization (Phan, Nguyen, & Horiguchi, 2008), and identification of sentiments in product reviews (Dave, Lawrence, & Pennock, 2003; Hu & Liu, 2004; Pang & Lee, 2008; Popescu & Etzioni, 2007). In the career domain, one could consider using written performance appraisals to predict dichotomously coded promotion decisions for a particular time period. For a fuller discussion and tutorial on text classification we refer the reader to Kobayashi et al. (2018a).

Neural Word Embeddings

Word embedding involves the mapping of a vocabulary of words to vectors of real numbers, this with the purpose of looking at the similarity and dissimilarity of the word vectors. Similarity and/or dissimilarity between vector representations can help gain intuition in the relation between words, for example "What are the top 5 most similar words to career satisfaction." Furthermore, vector representations allow us to perform certain arithmetic operations with an equal linguistic meaning, for example "Job" – "Pay" might result in a vector representation closest to "Voluntary work" (note that this is but an example).

A central paradigm for deriving these word vector representations is the distributional assumption on which it is based. First proposed by Harris in 1970, the distributional assumption states that words in similar contexts have similar meaning. For example, we could derive the meaning of the knowledge, skill, and abilities (KSAs) from certain occupations by looking at the context in which the KSAs occur. Thus, occupations related to "Computer Science," will then be describing a context containing KSAs such as; "data structures and algorithms," "optimization," "discrete mathematics," and so on. However, in order to fully capture the semantics of, for example, "Computer Science," domain-specific documents fully capturing the terminology describing the phenomenon "Computer Science" has to be eminent.

There are a variety of different methods to derive the word representations. Levy, Goldberg, and Dagan (2015) compared matrix-based algorithms, such as positive pointwise mutual information (PPMI) and singular value decomposition (SVD) to neural methods such as skip-gram with negative sampling (SGNS) and GloVe. Results found mostly local or insignificant performance differences between the methods, with no global advantage to any single approach over the others (Levy et al., 2015). For a more comprehensive overview and practical recommendations towards the implementation of the right method we refer the reader to Levy et al. (2015).

Validation and Postprocessing

The postprocessing step may involve domain experts to assist in determining how the output of the models can be used to improve existing processes, theory, and/or frame-

works. Two major issues are usually addressed here. The first is to find out whether the extracted patterns are real and not just random occurrences due to the sheer size of the data (e.g. by applying Bonferroni's principle). The second is, as with all empirical research, whether data and results are valid. Establishing validity (e.g. content, construct, internal, and external validity), and (therewith) the credibility of the output of TM models is particularly important for TM to gain legitimacy in careers research. It is important to note here that it is not the TM procedures that need to be validated but the output (in the same manner that we do not validate Factor Analysis), for example, the predictions of a TM based classifier.

Prior to being applied to support decision making and knowledge generation, the validity of TM based findings will need to be established. When TM is used to identify and operationalize key careers constructs, using different forms of data triangulation will help generate content and construct validity evidence. For example, in our job analysis example of TM application, which follows below, we enlisted the help of job analysts and subject matter experts in evaluating the output of the TM of vacancy texts. In other cases, TM outcomes could be compared to survey data, as was the case in the study on the role of personality in language use (Yarkoni, 2010). More generally, TM based models will require a comparative evaluation in which (part of) the TM output is correlated with independent or external data sources or other "standards" (such as the aforementioned survey or expert data). Though it is easy to view TM as a mechanistic means of extracting information from data, the input of domain/subject matter experts is critically important.

A straightforward practice for content validation is to have independent experts validate TM output. For example, in text classification, subject matter experts (SMEs) may be consulted from time to time to assess whether the resulting classifications of text are correct or not. A high agreement between the experts and the model provides an indication of the content-related validity of the model. The agreement is usually quantified using measures such as the Cohen's kappa or intra-class correlation coefficient.

Another way to validate TM output is through replication, data triangulation, and/or through an indirect inferential routing (Binning & Barrett, 1989). The standard can be established by obtaining external data using accepted measures or instruments that may provide theory based operationalizations that should or should not be correlated to the model. Such correlations give an indication of construct validity. For example, to validate experience requirements extracted from job vacancies, one can administer questionnaires to job incumbents asking them about their experience. Validity is then ascertained through the correlation between both operationalizations. This can be replicated on various types of text to assess if the TM model consistently generates valid experience requirements for a particular occupation. In theory, one could even compute full multi-trait multi-method correlation matrices (Campbell & Fiske, 1959) to compare the measurements obtained from TM with established instruments, although in practice it may be difficult to obtain the fully crossed dataset that it requires.

As with the statistical analyses that are (more) commonly applied in career research, text mining procedures in and of themselves cannot support causal inference (i.e. internal validity) unless the study design is such that, next to association, temporal precedence and isolation are also established. Given that many text mining applications rely on "data exhaust" (i.e. data that were not purposively collected for investigating the research questions at hand) and between-subjects designs, this is often difficult if not impossible to achieve. Despite these constraints, inferences pertaining to internal validity may be strengthened by collecting multiple waves of data and seeking to establish that changes in (a set of) TM based independent variable(s) are predictably related to changes in some hypothesized dependent variable(s) over time. Furthermore, although the ideal of randomized allocation of subjects to (pseudo-)experimental conditions may not always be viable, the sheer size of the data may likely yield sufficient statistical power to include a large number of theory based (Bernerth & Aguinis, 2016) control variables and/or to leverage a propensity score matching approach. Finally, as to external validity researchers need to be cognizant of the fact that although a given textual dataset may be qualified as being "big" it may still represent a non-randomly selected sample from a (much) larger population to which one wishes to generalize. The fact that Amazon ceased its résumé screening program, could be argued to have resulted from the fact that the (what turned out to be a gender biased) algorithm, was trained on an unrepresentative sample of data that was dominated by successful males. In our own work on vacancy mining we have also struggled with defining what our population is comprised of in the first place. Should the unit of analysis be vacancy, job, job candidate, or task? Clearly, whatever is decided upon has an important bearing on inferences pertaining to external validity. In sum, it is critical that the evaluation metrics developed and accepted within the data science tradition, must be augmented with the aforementioned validity types so as to facilitate adequate and accurate knowledge generation and decision making.

TEXT MINING EXAMPLE: SALARY PREDICTION FROM JOB VACANCIES

The purpose of this section is to illustrate how text mining may be applied in practice. The example uses job vacancies as a data source. The objective is to create a model that can be used to predict the salary associated with a particular occupation from job descriptions in job vacancies. The results could be used to shed light on the pay dynamics across jobs and may answer the question of what makes a particular job pay more as compared with other jobs. Although, there are many factors that affect pay (e.g. discrimination, geographic immobility, supply and demand of labor, etc.), here we examined how text mining could identify those antecedents that derive from the nature of the job. Also, based on the results we briefly discussed the issue of how skill requirements influence pay.

Salary Prediction from Job Vacancies

Every day, thousands of vacancies are posted to job boards and employment websites across the world. These vacancies are rapidly becoming a valuable source of job information. Most vacancies contain both worker-oriented (e.g. abilities, skills, knowledge, etc.) and job-oriented (e.g. work activities) information (cf. Peterson et al., 2001). Job analysts and labor organizations are looking for ways on how to use this abundant source of data to answer questions about jobs including evolution of skill demand, emergence of new jobs, and job task analysis. From the career research perspective, TM of vacancies can be to support career counseling and educational choices of students (Messum et al., 2017) and for quantifying the readiness of employees with respect to new work paradigms (Fareri et al., 2020).

A crucial step in using these vacancies is to be able to analyze their textual contents. For example, jobs may be grouped according to skill requirements or activities performed in the job. Also, job categories may reflect differences in salary. Classifying vacancies would give us better understanding of the job demand in each industry and may form the basis of further analysis such as the investigation of which skills are better compensated in the labor market.

Although there are already existing taxonomies of jobs and their associated pay systems, salary differences are still present even within the same job category and likely to change over time. Some researchers ascribe these differences to gender and/or salary negotiation skill (Säve-Söderbergh, 2019; Voigt & Ruppert, 2019). While other researchers point out other factors including artificial barriers (e.g. union and government restraints) and investment in human capital. In our small example, we will try to elucidate such salary differences just by analyzing job vacancies. Guided by the expansive literature on wage differentials, we attempt to identify some underlying factors influencing salary differences solely from analyzing vacancies. This approach applies a TM technique, specifically supervised LDA, to automatically extract topics from these vacancies and use these topics to create a predictive model of salary. The approach here is exploratory in nature but can be used as a basis for subsequent hypothesis testing. The following steps, which are explained in greater detail below, were followed: (1) Preprocess and transform the textual content of job vacancies, (2) extract topical patterns and build the model for salary prediction, and lastly, (3) evaluate the performance of the resulting model. This evaluation will be done by running the model on test data (i.e. data not used in building the model). Once we have determined that the model yields a reliable (consistent) prediction of salary, we can then develop strategies to establish evidence for content, construct, internal, and external validity, along the ways that were outlined earlier, although clearly some outcome and design would need to be defined for the establishment of internal validity. Although this process is unwieldy, and perhaps even beyond the scope of a single empirical study, ultimately, the insights we derive from examining how predictions are obtained and how they relate to other constructs may improve our understanding regarding pay differences in jobs and may form the basis for succeeding analyses.

Table 7.1 An example job vacancy text prior to and after text preprocessing

Original text	Resulting cleaned text
Apply now Castles Solicitors are looking for a part / full time Legal Secretary / Admin Assistant to join their team in Hurstpierpoint, working across multiple departments (flexible working hours available ideally 20+ hours a week) Essential skills: Microsoft and excel competentAbility to communicate with clients via email and telephoneOrganisation / good time managementAbility to work independently and as part of a teamGCSE Maths and English C or above Desirable but not essential as full training will be given: Legal background / qualificationsPrevious office experienceExperience with case management systems Role: Communicating with clients and creating appointments / general diary managementOpening and closing fileGeneral correspondence with clients and third partiesDrafting legal documentsTyping from dictationGeneral admin roles	apply now castle solicitor look part full time legal secretary admin assistant join team hurstpierpoint work across multiple department flexible work hour available ideally hour week essential skill microsoft excel competent ability communicate client via email telephone organisation good time management ability work independently part team gcse math english desirable essential full train will give legal background qualification previous office experience experience case management system role communicate client create appointment general diary management open close file general correspondence client 3 party draft legal document type dictation general admin role

Data

The data consisted of 50,000 job vacancies posted to various job boards in the United Kingdom. The automatic prediction of salary from job vacancy descriptions is expected to shed light on how salary differs across job groups. Furthermore, by analyzing the content of vacancies it should be possible to provide career options on jobs that pay better and a more detailed quantification of the skill–salary relationship, that is, how each skill is valued in the labor market.

Preprocessing

The first step in the process is text preprocessing. First, we extracted relevant content from text. Each vacancy is in HTML format, hence, HTML and other formatting tags (e.g. tabs, new line, and long whitespaces) were removed. Moreover, numbers, punctuation marks and stop words (Dolamic & Savoy, 2010; Fox, 1989) for the English language were deleted. Upper case letters were converted to lower case and finally all terms containing only two characters were removed. Extra whitespace and whitespaces at the beginning and end of the description were trimmed out. The extra whitespaces were the result of removing characters. An example job description and the resulting text after applying the preceding text cleaning procedures is shown in Table 7.1.

After text preprocessing, each text is then transformed into a vector resulting in a document-by-term matrix (DTM) representing the entire corpus. The columns of the DTM are the distinct terms occurring in the corpus and the entries are raw term frequencies. Since we want to run a supervised topic model, we need an outcome variable, in this case the outcome variable is annual salary, mentioned in the vacancies. For this, we developed a parser that automatically extracts the salaries. There

are a few nuances that we needed to deal with in order to extract the salaries, one is some salaries are provided on an hourly and monthly rate. Another is, some vacancies provide a range rather than a single figure for the salary. In order to harmonize the different salary rates, we converted all hourly and monthly salaries to annual salaries. For the hourly salary, we multiplied the hourly rate by 1538 which is the average hours per year that a full-time employee in the UK will spend working. In the same manner, we multiplied reported monthly salary by 12. For salary ranges, we computed mean salary or the middle of the range as an estimate of the salary and multiplied by an appropriate number to convert them to annual salaries. We put all extracted and calculated salaries in a column vector. Finally, we merged the column containing the salary of each job with the DTM.

Supervised LDA

For the supervised LDA part, we extracted one hundred topics. The choice of 100 topics is default in many other studies (Airoldi & Bischof, 2016; Wallach et al., 2009). However, the choice for the number of topics to retain can be evidenced by examining the prediction accuracy of models with varying numbers of topics. Since the purpose here is to illustrate rather than to optimize accuracy, the choice for 100 topics suffices in this case. The outcome of our supervised LDA here is a predictive model that is used for predicting salary (similar to a linear regression model). Aside from predicting salary, we can also examine the topics that were created. For purposes of illustration, we show nine topics in Table 7.2. The rest of the topics can be obtained upon request from the first author. In principal components analysis (PCA) in quantitative research, factor loadings are examined to interpret PC dimensions, in topic modeling, a topic can be interpreted by inspecting the top words, that is words that have the highest probabilities of belonging to the topic. The content of the vacancies is summarized by the topics such as shown in Table 7.2. Some topics are indicative of the job requirements written in vacancies. Finally, Table 7.3 shows the salary ranges and sample job titles associated to topics 3, 9, 32, 33, and 76.

Predictive performance is measured using the mean squared errors which is computed by the following formula:

$$\text{mean squared error} = \frac{1}{n}\sum_{i=1}^{n}\left(observed_i - predicted_i\right)^2$$

where n is the size of the corpus. The performance was evaluated using 10-fold cross validation Kohavi (1995).

Table 7.2 Sample topics constructed using supervised topic model

Topic 8	**Topic 19**	**Topic 20**
food	member	software
chef	ability	developer
restaurant	relevant	technology
kitchen	communicate	test
hotel	effectively	web
hospitality	communication	java
guest	check	net
cater	write	sql
head	responsibility	script
fresh	clear	end
Topic 25	**Topic 27**	**Topic 29**
digital	maintenance	safety
medium	machine	health
campaign	installation	energy
social	system	requirement
content	carry	responsibility
brand	equipment	legislation
channel	tool	environmental
agency	gas	assessment
online	repair	current
strategy	test	qualification
Topic 40	**Topic 61**	**Topic 86**
datum	care	engineer
analyst	people	manufacture
analysis	worker	maintenance
report	social	production
insight	child	mechanical
analyse	young	electrical
model	health	equipment
analytic	home	technical
analytics	community	plant
use	life	electronic

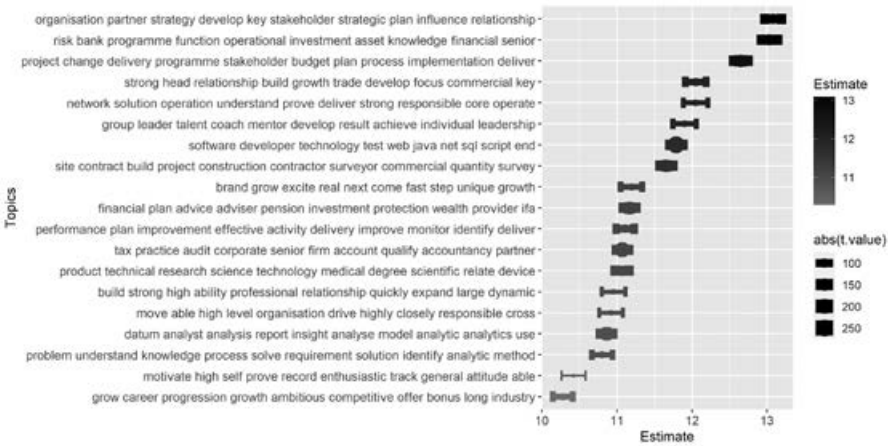
Results

The 10-fold cross validation yielded an average mean squared error of 0.16 which denotes good performance. For the postprocessing, and to establish construct validity, the model is applied to predict the salary in the vacancies in the test data and it yielded a mean squared error of 0.23. We also used the output of the model to examine which topics are associated with high and low salaries. Figure 7.2 shows the highest five topics and lowest five topics as they relate to salary. The figure shows that jobs about strategy planning, banking jobs, and jobs located in big cities (e.g. London and Birmingham) tend to pay more than cleaning and call center jobs. Figure 7.3 shows the ordering of skills and knowledge as they relate to salaries. As can be

Table 7.3 Job titles, salary ranges associated with topics 3, 9, 32, and 76

Topics	Terms	Average salary	Top job titles
3	organisation partner strategy develop key stakeholder strategic plan influence relationship	50,000–115,000	"Head of Communications and Engagement," "Head of Internal Communications & Engagement," "Head of Partnerships and Performance," "Stakeholder Engagement Strategy Manager Commercial Excellence," "Chief Product Officer"
9	risk bank programme function operational investment asset knowledge financial senior	90,000–132,000	"Operational Risk Manager," "Senior Investment Risk Manager," "Financial Intelligence Specialist," "Treasury Analyst, Liquidity Management, ALM, Banking," "Brexit Risk Manager – Insurance Delegated Authority Background," "Chief Operation Officer, FX trading"
32	clean area facility duty use keep general part equipment site	16,800–26,000	"School Cleaner," "Housekeeping Assistant/Laundry Assistant," "Cleaning Operatives/Cleaners Night – Gatwick Airport," "Refuse/Recycling Loader," "Aircraft Cleaning Operatives- Heathrow Airport," "Night Cleaning Manager"
33	detail ability attention communication deadline pressure write organisational verbal able	25,000–50,000	"Litigation Support Specialist – Northampton," "Trade and Transaction Reporting Officer," "Recruitment Resourcer – Admin," "Sales Administrator," "Marketing CRM Assistant," "Desktop Publishing Specialist," "IT Project Support Analyst," "Administrator," "Temporary Sales Administrator," "Sales Order Processor," "Immigration Solicitor," "Costs Draftsman"
76	call centre free park yorkshire advisor contact inbound outbound position	22,500–78,000	"Call Centre Advisor," "Your 1st Call Centre Customer Service Advisor Job," "Receptionist," "Assistant Buyer/Purchasing Assistant," "Retail/Shop Supervisor," "Digital Marketing Executive," "Customer Services Adviser – Investment and Financial Services," "Telesales Executive," "Collections Advisor"

Figure 7.2 *Top five topics associated to higher pay (first five) and topics associated to least pay (last five)*



Note: The x-axis represents the coefficients of the predictive model; the higher the coefficients the higher the contribution to salary. The error bars represent the standard errors of the coefficients.

Figure 7.3 *Knowledge and skills associated with higher pay*

seen, knowledge on finance, software development and data analysis command high salaries in the labor market. Moreover, organizing, leadership, and problem-solving skills are also highly valued. Other employee characteristics that also appear to incur higher pay are being motivated and ambitious. It is surprising to see that communication skills does not figure in the list of high paying skills. One explanation is this skill is commonly required and that it is usually stated in low paying jobs such as in

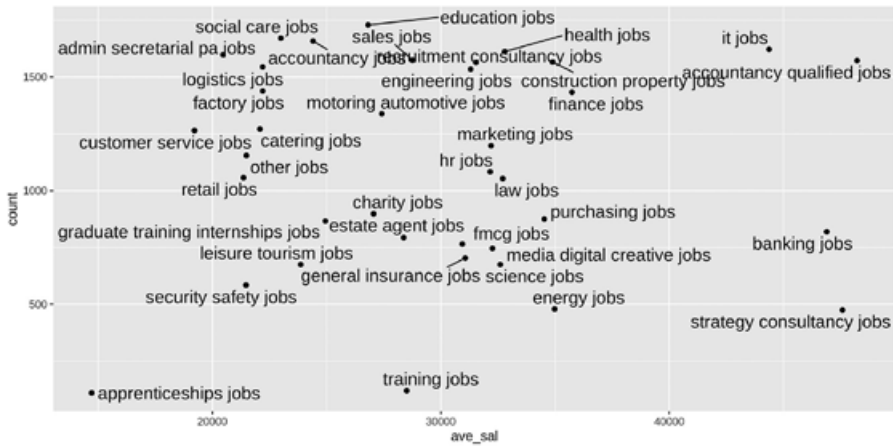
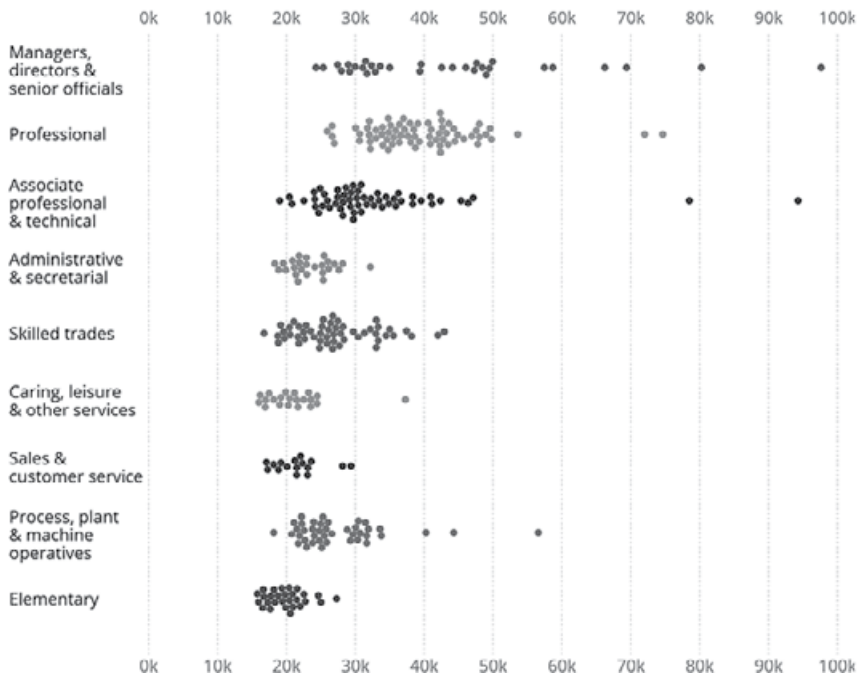


Figure 7.4a Job categories according to salary and availability

call centers or administrative support. We investigated further by analyzing which job categories tend to pay higher, Figure 7.4a shows a scatter plot of jobs according to average salary and availability. It is obvious that some high paying jobs are not necessarily in high demand, this can be explained by the fact that these jobs require higher ability and more experience (more investment in human capital). Another explanation is that the productivity of workers in these job categories contribute more to the revenue of the companies (banking, strategy, and consultancy jobs). Two job categories seem to offer very good opportunity/outlook because of their high demand and pay, these are IT and accountancy qualified jobs. Professional jobs lie somewhere above the median salary and jobs in retail, admin, social care, customer service or related to leisure are the least paying jobs, with the exception of apprenticeship jobs which usually are unpaid jobs. We compared our results with the data from the National Statistics Office of UK (Smith, 2019) (Figure 7.4b) and we find similarity in terms of the ranking of job categories with respect to salary. The similarity provides an evidence of the content validity of the information extracted from job vacancies.

This example showed the feasibility of developing a prediction model for the automatic prediction of salary from text. The resulting classifier is scalable as it can predict salaries on thousands of vacancies which would be laborious and time consuming when done manually. Moreover, the model also sheds light on what determines salaries in jobs which can be difficult to determine especially in the case of new jobs.



Source: Reprinted from Smith (2019). Public sector information licensed under the Open Government Licence v3.0.

Figure 7.4b Plot of salaries according to job categories

Other Potential Applications

In the preceding section we applied a text mining technique (e.g. supervised topic modeling) on a problem not only of potential interest to career researchers but also to labor economists, and education researchers. Here we explore other potential avenues for future research at the interface between text mining and careers research by suggesting some ideas how existing problems in career research can be reconceptualized as a text mining problem. Our recommendations are purely speculative at the moment and are aimed to excite and provide inspiration to career researchers to incorporate text in their analysis as well as to encourage them to collaborate with text mining experts.

Studies that attempt to analyze how individual attributes relate to career outcomes may benefit from text mining. Specifically TM may provide highly scalable and unobtrusive means of assessing those psychological constructs that careers researchers often hypothesize to predict career outcomes. Not only would this approach prevent burdening respondents with lengthy surveys, it also has the potential to

generate data for much larger samples than we are typically accustomed to. One may build a simple score-based system that measures how each word is related to a certain a career concept or dimension using frequency analysis or word-embedding approaches (Shen, Brdiczka, & Liu, 2013; Yarkoni, 2010). Indeed, results from TM can be compared to Likert-scale questionnaires and when the results are properly validated, they may even be used in situations where it is not feasible to gather participants to complete survey measures, for example in observational studies or in the use of secondary text (web blogs, speeches, essays, etc.).

Another potential application is in helping students shape career interests and paths that may lead to satisfaction/well-being. For example, in conceptualizing personal goals, students may be encouraged to write about their career goals in narrative form. By running topic modeling, we may be able to surmise pivotal goal components that relate to career satisfaction. Also, we may build classification models to automatically sort goals into choice or performance goals. The classification is trained by preparing a training data where subject matter experts (SMEs) annotate the goal narratives by manually classifying the mentioned goals into the two categories. Using the annotated narratives, a classification model can be trained and subsequently applied to other goal narratives.

Text mining can also be used in the analysis of mentoring relationships such as in coaching evaluation. Analyzing the written evaluation of or testimonials about career development coaches may be used to analyze which criteria are important for training and development. By analyzing coaches' evaluation using latent semantic analysis (LSA) we may be able to untangle how coaches differ in their evaluation based on length of experience and expertise (Theeboom et al., 2017). This can be done by relating the results of LSA to training outcomes.

Perhaps, one big advantage of text mining in today's research landscape is its ability to analyze big text data that transcend organization and geographical boundaries. Nowadays, online text data about almost anything are readily available and can be collected via an application programing interface (API) or web crawling. Also, these text data can be routinely analyzed even in the absence of a solid theoretical model. A newer approach is to run text analysis on a specific corpus, evaluate whether the patterns extracted are of practical/theoretical significance, before digging deeper in the hope that the extracted information can be used as a basis for building conceptual models (i.e. as is done in grounded theory approach). This approach is particularly useful to study new phenomena in careers research. For example, in the study of new forms of organizing problems, Karanović, Berends, and Engel (2020) applied structural topic modeling to analyze how workers in the platform economy engage with novel forms of organizing across regulatory structures. By analyzing posts in fora set up by the workers themselves, the authors found that workers respond differently to organizing solutions imposed by these platforms. Hence, this may reveal a new dimension of worker–employer relationship in platforms where this relationship is fuzzy. Note that using traditional methods in careers research to investigate this phenomenon may be challenging because it involves non-traditional workers who are dispersed across the globe.

Table 7.4 Potential applications of text mining in career research

Career topics	Text source	Text mining methods	Validation	Sample studies
Study of psychological constructs as they relate to career outcomes	Responses in open-ended questions	Frequency analysis, word embedding	Matched to validated Likert-style questionnaires	Shen et al. (2013); Yarkoni (2010)
Training and development	Coaching/trainer evaluation	Latent semantic analysis	Validate through subject matter experts and data triangulation. Also through predictive validity	Theeboom et al. (2017)
New employment relationships	Posts in online forum, web blogs, and twitter data	Structured topic modeling	Subject matter experts	Karanović et al. (2020)
Job turnover intention	Articles and exit interview transcripts	Text classification and frequency analysis	Matched to turnover decisions; convergent validity by investigating its relationship to organizational commitment, job satisfaction, and organizational justice	Barrick & Zimmerman (2005); Frederiksen (2017); Lee, Kim, & Mun (2019)
Work/non-work conflict	Diary; interview transcript; self-evaluation	Longitudinal topic modeling (e.g. dynamic topic models)	Convergent validity as well as predictive validity by relating it to family, physical and psychological health outcomes	Dettmers (2017); Judge, Van Vianen, & De Pater (2004)
Expatriate assignments	Expatriates narratives from semi-structured interviews	Topic modeling and text classification	Subject matter experts (e.g. expatriates coaches)	Salomaa & Makela (2017); Wechtler et al. (2018)

Finally, text mining is not limited to the above-mentioned topics or applications, it can be used to study well-established career topics such as job turnover (Frederiksen, 2017), job burnout (Lizano & Barak, 2015; Reid, Short, & McKenny, 2017), work/non-work conflict (Dettmers, 2017), and expatriation (Wechtler, Koveshnikov, & Tienari, 2018). As long as there are free texts one can get hold of, the limits of investigation lie on the researchers' creativity and imagination. In order to proceed with the analysis, our recommendation is to first investigate whether existing conceptual models are available to guide the text mining process. A model can be used to judge the validity of the extracted patterns. However, the models should not limit the inquiry as potentially interesting patterns may emerge. New interesting patterns may be used to improve or even to refute the model. The researcher may subject the

patterns to both data triangulation and validation (e.g. use of subject matter experts, convergent validity and predictive validity) as this will enhance the credibility of the newly discovered concept. Table 7.4 summarizes how existing and new career concepts can be analyzed through text mining.

CONCLUSION

The proliferation of text data about peoples' career and the development of text analytical techniques hold great promise in accelerating and augmenting career research. However, up to this time, most career studies seldom utilize unstructured text as data, and if used, text analysis is usually limited to counting word frequencies. Text mining enables career researchers to augment their text analytical toolkit with powerful analytical techniques from machine learning and statistics. As illustrated in this chapter, text mining is a process that broadly consists of three iterative phases: text preprocessing, the application of TM operations, and finally postprocessing. We hope to have provided readers with: (i) a general understanding of how text is transformed to a form amenable to the application of analytical techniques, (ii) the ability to identify appropriate TM operations to address a research problem, and (iii) tactics to validate the results. Finally, TM derived insights may be further investigated through hypothesis testing in order to inform theory. This perhaps is the real crux why career researchers may be interested in using text mining in their investigations. Although, this chapter is largely written to instruct, it is our hope that through this chapter more career researchers will augment their analytical toolkit with TM, and that this in turn will lead to a better understanding of how careers develop.

NOTES

2. This work was supported in part by the European Commission through the Marie-Curie Initial Training Network EDUWORKS (Grant PITN-GA-2013-608311).
Parts of this chapter were adopted with permission from: Kobayashi et al. (2018a), licenced under CC-BY-NC 4.0 and Kobayashi et al. (2018b), licenced under CC-BY-NC 4.0.
Correspondence concerning this chapter should be addressed to Stefan T. Mol, Leadership and Management Section, Amsterdam Business School, University of Amsterdam, Valckenierstraat 59, 1018 XE, Amsterdam, the Netherlands; email: s.t.mol@uva.nl; Tel.: +31-(0)20-525 5490.
3. Raw TF refers to the actual frequency of the word in a document.

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 163–222). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_6

Airoldi, E. M., & Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, *111*(516), 1381–1403. https://doi.org/10.1080/01621459.2015.1051182

Alpaydin, E. (2014). *Introduction to Machine Learning*. MIT Press.

Barrick, M. R., & Zimmerman, R. D. (2005). Reducing voluntary, avoidable turnover through selection. *Journal of Applied Psychology*, *90*(1), 159–166. https://doi.org/10.1037/0021-9010.90.1.159

Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, *69*(1), 229–283. https://doi.org/10.1111/peps.12103

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 245–250. http://dl.acm.org/citation.cfm?id=502546

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478–494. https://doi.org/10.1037/0021-9010.74.3.478

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Conrad, J. G., Al-Kofahi, K., Zhao, Y., & Karypis, G. (2005). Effective document clustering for large heterogeneous law firm collections. *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, 177–187. https://doi.org/10.1145/1165485.1165513

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, 519–528. https://doi.org/10.1145/775152.775226

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*, *1644*(1), 97–104. https://doi.org/10.1063/1.4907823

Derpanis, K. G. (2006). *K-Means Clustering*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.217.5155

Dettmers, J. (2017). How extended work availability affects well-being: The mediating roles of psychological detachment and work-family-conflict. *Work & Stress*, *31*(1), 24–41. https://doi.org/10.1080/02678373.2017.1298164

Dolamic, L., & Savoy, J. (2010). When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*, *61*(1), 200–203. https://doi.org/10.1002/asi.21186

Fareri, S., Fantoni, G., Chiarello, F., Coli, E., & Binda, A. (2020). Estimating Industry 4.0 impact on job profiles and skills using text mining. *Computers in Industry*, *118*, 103222. https://doi.org/10.1016/j.compind.2020.103222

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291–314.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Fox, C. (1989). A stop list for general text. *SIGIR Forum*, *24*(1–2), 19–21. https://doi.org/10.1145/378881.378888

Frakes, W. B., & Baeza-Yates, R. (Eds.) (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice Hall.

Frederiksen, A. (2017). Job satisfaction and employee turnover: A firm-level perspective. *German Journal of Human Resource Management*, *31*(2), 132–161.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157–1182.

Harish, B. S., Guru, D. S., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. *International Journal of Computer Applications IJCA*, *RTIPPR*(2), 110–119.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. In J. Euzenat & J. Domingue (Eds.), *Artificial Intelligence: Methodology, Systems, and Applications* (Vol. 4183, pp. 77–86). Springer. https://doi.org/10.1007/11861461_10

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. http://dl.acm.org/citation.cfm?id=1014073

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323.

Jolliffe, I. (2005). *Principal Component Analysis*. Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1002/0470013192.bsa501/full

Judge, T. A., Van Vianen, A. E. M., & De Pater, I. E. (2004). Emotional stability, core self-evaluations, and job outcomes: A review of the evidence and an agenda for future research. *Human Performance*, *17*(3), 325–346.

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer Science & Business Media.

Karanović, J., Berends, H., & Engel, Y. (2020). Regulated dependence: Platform workers' responses to new forms of organizing. *Journal of Management Studies*, in press. https://doi.org/10.1111/joms.12577

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018a). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*, *21*(3), 766–799. https://doi.org/10.1177/1094428117719322

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018b). Text mining in organizational research. *Organizational Research Methods*, *21*(3), 733–765. https://doi.org/10.1177/1094428117722619

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, pp. 1137–1145). Retrieved from http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf

Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 721–735. https://doi.org/10.1109/TPAMI.2008.110

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lee, N., Kim, J.-H., & Mun, H.-J. (2019). Exploration of Emotional Labor Research Trends in Korea through Keyword Network Analysis. *Journal of Convergence for Information Technology*, *9*(3), 68–74. https://doi.org/10.22156/CS4SMB.2019.9.3.068

Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An empirical comparison of four text mining methods. *2010 43rd Hawaii International Conference on System Sciences (HICSS)*, 1–10. https://doi.org/10.1109/HICSS.2010.48

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. https://doi.org/10.1162/tacl_a_00134

Lewis, D. D. (1992). *Representation and learning in information retrieval* [University of Massachusetts]. http://ciir.cs.umass.edu/pubfiles/UM-CS-1991-093.pdf

Lizano, E. L., & Barak, M. M. (2015). Job burnout and affective wellbeing: A longitudinal study of burnout and job satisfaction among public child welfare workers. *Children and Youth Services Review*, *55*, 18–28.

McAuliffe, J. D., & Blei, D. M. (2008). Supervised topic models. *Advances in Neural Information Processing Systems*, 121–128.

Messum, D., Wilkes, L., Peters, K., & Jackson, D. (2017). Content analysis of vacancy advertisements for employability skills: Challenges and opportunities for informing curriculum development. *Journal of Teaching and Learning for Graduate Employability*, *7*(1), 72–86. https://doi.org/10.21153/jtlge2016vol7no1art582

Montanelli Jr, R. G., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika*, *41*(3), 341–348.

Osinski, S., & Weiss, D. (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, *20*(3), 48–54. https://doi.org/10.1109/MIS.2005.38

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., Campion, M. A., Mayfield, M. S., Morgeson, F. P., Pearlman, K., Gowing, M. K., Lancaster, A. R., Silver, M. B., & Dye, D. M. (2001). Understanding work using the occupational information network (o*net): Implications for practice and research. *Personnel Psychology*, *54*(2), 451–492. https://doi.org/10.1111/j.1744-570.2001.tb00100.x

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceedings of the 17th International Conference on World Wide Web*, 91–100. https://doi.org/10.1145/1367497.1367510

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In A. Kao & S. R. Poteet (Eds.), *Natural Language Processing and Text Mining* (pp. 9–28). Springer London. https://doi.org/10.1007/978-1-84628-754-1_2

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 569–577. http://dl.acm.org/citation.cfm?id=1401960

Reid, S., Short, J. C., & McKenny, A. (2017). Tell me how you feel: A content analytic approach to measuring burnout. *Academy of Management Proceedings*, *2017*, 14641.

Salomaa, R., & Makela, L. (2017). Coaching for career capital development: A study of expatriates' narratives. *International Journal of Evidence Based Coaching and Mentoring*, *15*(1), 114–132.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Säve-Söderbergh, J. (2019). Gender gaps in salary negotiations: Salary requests and starting salaries in the field. *Journal of Economic Behavior & Organization*, *161*, 35–51.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. *Proceedings of the Sixteenth International Conference on Machine Learning*, 379–388. http://dl.acm.org/citation.cfm?id=645528.657484

Shen, J., Brdiczka, O., & Liu, J. (2013). Understanding email writers: Personality prediction from email messages. *User Modeling, Adaptation, and Personalization*, 318–330. https://doi.org/10.1007/978-3-642-38844-6_29

Smith, R. (2019). *Employee Earnings in the UK: 2019*. Office for National Statistics. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2019#employee-earnings-data

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Analysis and Applications*, *8*(1–2), 199–209. https://doi.org/10.1007/s10044-005-0256-3

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining*, *400*, 525–526. https://wwws.cs.umn.edu/tech_reports_upload/tr2000/00-034.ps

Theeboom, T., Van Vianen, A. E. M., Beersma, B., Zwitser, R., & Kobayashi, V. (2017). A practitioner's perspective on coaching effectiveness. In L. Nota & S. Soresi (Eds.), *Counseling and Coaching in Times of Crisis and Transitions: From Research to Practice* (pp. 61–78). Routledge.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129–140. https://doi.org/10.1023/B:ETIN.0000047476.05912.3d

Vempala, S. S. (2005). *The Random Projection Method* (Vol. 65). American Mathematical Society.

Voigt, M., & Ruppert, A. (2019). Salary negotiations: highlights and surprises from 10 years of research. *International Conference on Gender Research*, 639–645.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112. http://dl.acm.org/citation.cfm?id=1553515

Wechtler, H., Koveshnikov, A., & Tienari, J. (2018). The spouse as "the forgotten other": A qualitative content analysis of expatriation literature. *Academy of Management Proceedings*, *2018*(1), 18361. https://doi.org/10.5465/AMBPP.2018.18361abstract

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of Predictive Text Mining* (2nd ed.). Springer-Verlag London. https://doi.org/10.1007/978-0-387-34555-0_2

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*, 412–420. http://dl.acm.org/citation.cfm?id=645526.657137

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Youn, S., & McLeod, D. (2007). A comparative study for email classification. In K. Elleithy (Ed.), *Advances and Innovations in Systems, Computing Sciences and Software Engineering* (pp. 387–391). Springer. http://link.springer.com/chapter/10.1007/978-1-4020-6264-3_67

Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, *21*(8), 879–886. https://doi.org/10.1016/j.knosys.2008.03.044

Zhang, Y., Chen, M., & Liu, L. (2015). A review on text mining. *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 681–685. https://doi.org/10.1109/ICSESS.2015.7339149