Automatic identification of competences expected by employers with the use of exploratory text analysis

Paweł Lula[1], Renata Oczkowska[2], Sylwia Wiśniewska[3]

Abstract

Exploratory text analysis allows to identify semantic components present in processing documents. For every component it is possible to describe its character and to evaluate its importance. Using the approach presented above for automatic analysis of job offers it is possible do discover components which are common for all texts and to estimate their importance in every offer. Unfortunately, semantic components obtained with the help of text mining algorithms, usually do not reflect competences within the meaning of specialists from the HR area.

In the paper authors are going to present a method which will be able to identify in a set of job offers semantic components corresponding to professional, social, personal or managerial competences. Also the method of competence description and evaluation will be proposed. The computational model used for the analysis is composed of the two parts. The first is formed by the Latent Dirichlet Allocation Model and identifies latent semantic components. The second element of the model maps semantic components calculated by the *LDA* approach into the set of components related to employers' expectations towards candidates for employment. The proposed method was used for analysis of the corpus containing job offers related to the field of human resources management.

Keywords: *competences, the Latent Dirichlet Allocation Model, job offer exploratory analysis*
JEL Classification: J24, C81
DOI: 10.14659/SEMF.2018.01.26

1 Introduction

Continuous analysis of employers' expectations towards candidates for employment and competences of candidates who are actively looking for a new job can be helpful for all institutions that have impact on modern labour market, particularly for universities which are responsible for development of professional education essential for future employees.

The analysis of employers' expectations can be performed with the use the exploratory analysis methods (Manning and Schütze, 1999), (Salton et. al., 1975). Latent Dirichlet Allocation (*LDA*) method (Blei, 2003), (Alghamdi and Alfalqi, 2015) seems to be useful for identification of semantic components (so called *topics*) in a given set of offers.

[1] Corresponding author: Cracow University of Economics, Department of Computational Systems, 27 Rakowicka Street, 31-510 Kraków, Poland, pawel.lula@uek.krakow.pl
[2] Cracow University of Economics, Department of Human Resources Management, 27 Rakowicka Street, 31-510 Kraków, Poland, renata.oczkowska@uek.krakow.pl
[3] Cracow University of Economics, Department of Human Resources Management, 27 Rakowicka Street, 31-510 Kraków, Poland, sylwia.wisniewska@uek.krakow.pl

Unfortunately, components provided by the *LDA* algorithm not always directly represent features which are taken into account by employers and by people responsible for employment policy. Therefore it is necessary to map *LDA* components into the space of competences which is defined by a set of features significant for labour market researchers and analysts.

The method proposed in the paper assumes that components calculated by the use of the *LDA* method are transformed by the classification model with class overlapping which generates labels and evaluates competences' importance by calculation assignment coefficients for every competence. In our approach a classification model has been built in a supervised mode, with the use of a training set retrieved from the *pracuj.pl* web site.

In the second section of the paper some definitions of competences are discussed. Also a competence taxonomy is presented. Next, in the third section, a proposed method of competences' identification is proposed. Some exemplary analysis are shown in the fourth part of the paper. Final conclusions are formulated in the last part of the paper. All calculations presented in the article were performed in R language.

2 The essence of competences

Studies on the subject literature indicates that currently prevail two approaches, according to which the concept of employee competences is interpreted. According to the first approach, competences are seen as people's characteristics, which account for a basis of desired behaviour at work, allowing them to achieve intended results. Competences are therefore understood as the ability to implement specific patterns of behaviour. According to the second approach, competences are understood as characteristics of a professional position. This interpretation defines competences as the ability to effectively perform professional duties in compliance with standards established by the organisation or to achieve desired results (Whiddett and Hollyforde, 2003).

The first approach is presented by Whiddett and Hollyforde (2007). They define competences as "the behaviours that individuals demonstrate when undertaking job relevant tasks effectively within a given organisational context". According to the second approach, the concept of competence is defined by Wright et. al. (2003). In their opinion, competence is "the ability to perform activities within an occupation or function to the standards expected in employment".

Analysis of the subject literature point at the diversity of typologies of competences. Among the most popular is the classification of Filipowicz (2014). He distinguishes the following types of competences:

1. social – determining the quality performed tasks associated with contacts with people (e.g. commercial contacts). The level of these competences determines the effectiveness of communication, cooperation and influencing other people. Sample social competences include, among others: oral communication, written communication, teamwork, building of relations, sharing of knowledge and experience;
2. personal – related to performance of tasks by the employee, and their level affects the quality of the performed tasks. They also determine the adequacy in actions and the speed of their performance. Personal competences include, among others: innovativeness, entrepreneurship, flexibility, organization of own work, time management, problem-solving, and stress management;
3. managerial – involve human resource management, both with soft areas of management, work organisation, as well as with strategic aspects of management. The effectiveness of management of the organisation's employees is determined by the level of these competences. Exemplary managerial competences include, among others: analytic thinking, strategic thinking, motivating, delegation of tasks, team-building;
4. professional (specialist, technical) – concern specialist tasks set for particular groups of positions. They are often connected with specific scopes of knowledge or skills. The level of these competences is reflected in the effectiveness of implementation of tasks typical of a given profession, position or performed function. Examples of professional competences include: process management, project management, professional knowledge, the ability to use modern information technologies.

The classification of competences presented above will be used as the basis of the computational model described in the next section.

3 Identification of competences in textual job offers

Automatic mapping job offer content into the space defined by competences relevant to human resources management specialists and labour market researchers and analysts is the main aim of the study (Oczkowska et. al., 2017). The proposed model has compound character and is composed of two sub-models:

- the Latent Dirichlet Allocation model used as a filter for erasing irrelevant elements from documents and for documents mapping into the space of semantic components (*topics*). However, topics identified by the *LDA* algorithm usually are not consistent with competences which are studied in the area of human resources management,
- the classification model with overlapping classes which is responsible for mapping documents from the space defined by *LDA* topics to the space of competences.

The structure of the model is presented in the Fig. 1.

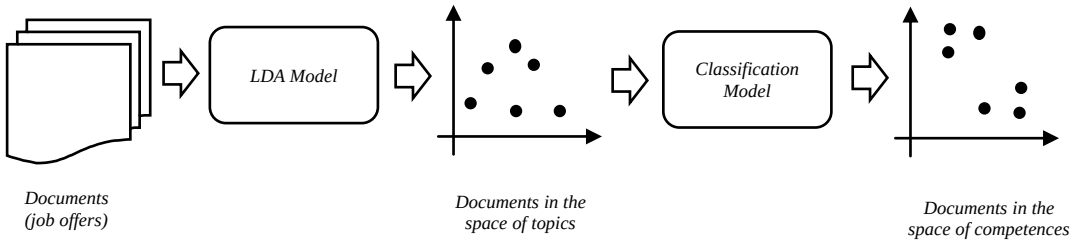

Fig. 1. The structure of the model mapping job offers into the space of competences.

3.1 The Latent Dirichlet Allocation Model

The process of analysis starts with the job offers processing with the *Latent Dirichlet Allocation* (*LDA*) model.

The *LDA* algorithm analyses the corpus which is composed of *LD* documents (it was assumed that one document from the corpus contains one job offer):

$$\boldsymbol{D} = [d_1 \quad \dots \quad d_{LD}]^T. \tag{1}$$

Words used in documents form the dictionary $\boldsymbol{V}$ containing *LV* terms:

$$\boldsymbol{V} = [v_1 \quad \dots \quad v_{LV}]^T \tag{2}$$

The *LDA* model identifies LT semantic components (topics) which reflect main issues appearing in the corpus. Topics constitute $\boldsymbol{T}$ structure:

$$\boldsymbol{T} = [t_1 \quad \dots \quad t_{LT}]^T \tag{3}$$

The results of the *LDA* models are presented as two matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. The $\boldsymbol{\Phi}$ matrix defines all topics identified by the *LDA* algorithm. Every topic is represented by the probability distribution over words taken from the dictionary $\mathbf{V}$. The $\boldsymbol{\Phi}$ matrix has a following form:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,LV} \\ \cdots & \cdots & \cdots \\ \phi_{LT,1} & \cdots & \phi_{LT,LV} \end{bmatrix} \tag{4}$$

where $\phi_{i,j}$ represents the probability of occurrence of the j-th word from the dictionary $\mathbf{V}$ in the i-th topic.

The $\mathbf{\Theta}$ matrix informs about the contribution of every topic to every document and has a following form:

$$\mathbf{\Theta} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,LT} \\ \cdots & \cdots & \cdots \\ \theta_{LD,1} & \cdots & \theta_{LD,LT} \end{bmatrix} \tag{5}$$

where $\theta_{i,j}$ indicates the probability of occurrence of the j-th topic in the i-th document.

3.2 The Classification Model

The classification model should map a given document represented as a point in the space of *LDA* topics into the space of competences. Taking into account that one job offer includes requirements related to many competences, the classification model should calculate assignment coefficients representing the contribution of every competence to a given document.

Establishing the set of competences which should be analysed by the model constitutes the first step in the process of classification model building. The set of competences includes LC elements and can be defined as:

$$\boldsymbol{C} = \{C_1, C_2, \ldots, C_{LC}\}. \tag{6}$$

In contrast to the *LDA* model, the classification model was estimated in the supervised mode with the use of LC text files, where file f_i contains exemplary sentences derived from job offers related to the competence C_i. For every competence, on the basis of a file with description of requirements related to this competence, the probability distribution over words has been calculated. The result forms the $\mathbf{\Gamma}$ matrix:

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{1,1} & \cdots & \gamma_{1,LV} \\ \cdots & \cdots & \cdots \\ \gamma_{LC,1} & \cdots & \gamma_{LC,LV} \end{bmatrix} \tag{7}$$

where $\gamma_{i,j}$ is the probability of occurring of the j-th word in the description of the i-th competence.

Next the similarity of every *LDA* topic to every competence was calculated. In the current study the cosine similarity was used.

$$\delta^*_{i,j} = \frac{\sum_{v=1}^{LV} \phi_{i,v}\gamma_{j,v}}{\sqrt{\sum_{v=1}^{LV} \phi_{i,v}^2}\sqrt{\sum_{v=1}^{LV} \gamma_{j,v}^2}} \tag{8}$$

Similarity coefficients form a matrix with LT rows and LC columns. Dividing every element of this matrix by the sum of row elements, the $\mathbf{\Delta}$ matrix was obtained:

$$\mathbf{\Delta} = \begin{bmatrix} \delta_{1,1} & \cdots & \delta_{1,LC} \\ \cdots & \cdots & \cdots \\ \delta_{LT,1} & \cdots & \delta_{LT,LC} \end{bmatrix} \tag{9}$$

where:

$$\delta_{i,j} = \frac{\delta^*_{i,j}}{\sum_{k=1}^{LC} \delta^*_{k,j}} \tag{10}$$

The $\delta_{i,j}$ value represents the distribution of the j-th competence in the i-th *LDA* topic.

Using the document representation in the topic space, the contribution of the j-th competence in the i-th job offer can be expressed as:

$$s_{i,j} = \sum_{k=1}^{LT} \theta_{i,k} \delta_{k,j} \tag{11}$$

Finally, the description of documents in the space of competences is defined as the **K** matrix:

$$\mathbf{K} = \begin{bmatrix} \kappa_{1,1} & \cdots & \kappa_{1,LC} \\ \cdots & \cdots & \cdots \\ \kappa_{LD,1} & \cdots & \kappa_{LD,LC} \end{bmatrix} \tag{12}$$

where element:

$$\kappa_{i,j} = \frac{s_{i,j}}{\sum_{k=1}^{LC} s_{i,k}} \tag{13}$$

defines the importance of the j-th competence in the i-th offer.

4 The analysis of the competence expectations based on the exploratory analysis of job offers related the HR management area

During the study, a set of job offers related to the human resources management area was analysed. All offers were retrieved from the *pracuj.pl* web portal. For offers retrieving the *rvest* package was used (Wickham, 2016). The corpus was prepared with the help of the *tm* package (Feinerer et. al., 2008).

The following set of competences was taken into account:

$$\boldsymbol{C} = \{M_1, P_1, P_2, P_3, S_1, S_2, T_0, T_1, T_2, T_3, T_4, T_5\} \tag{14}$$

where:

- M_1 – managerial competences,
- P_1 – personal competences (innovativeness, problem solving ability, dealing with stress, ...),
- P_2 – possession of a university diploma,
- P_3 – possession of a driving licence,
- S_1 – communication competences (oral and written, also in foreign languages),
- S_2 – social competences,
- T_0 – general competences and experience in HR area,

- T_1 – legal and organizational aspects of HR management,
- T_2 – competences in recruitment and training,
- T_3 – IT competences,
- T_4 – competences in management, accounting and logistics,
- T_5 – competences in sales and customer relationship management.

Next, with the use of randomly chosen offers prepared in Polish, the matrix **Γ** was calculated. As an example, in the Fig. 2 the distribution of the most important words for the T_1 competence was presented.

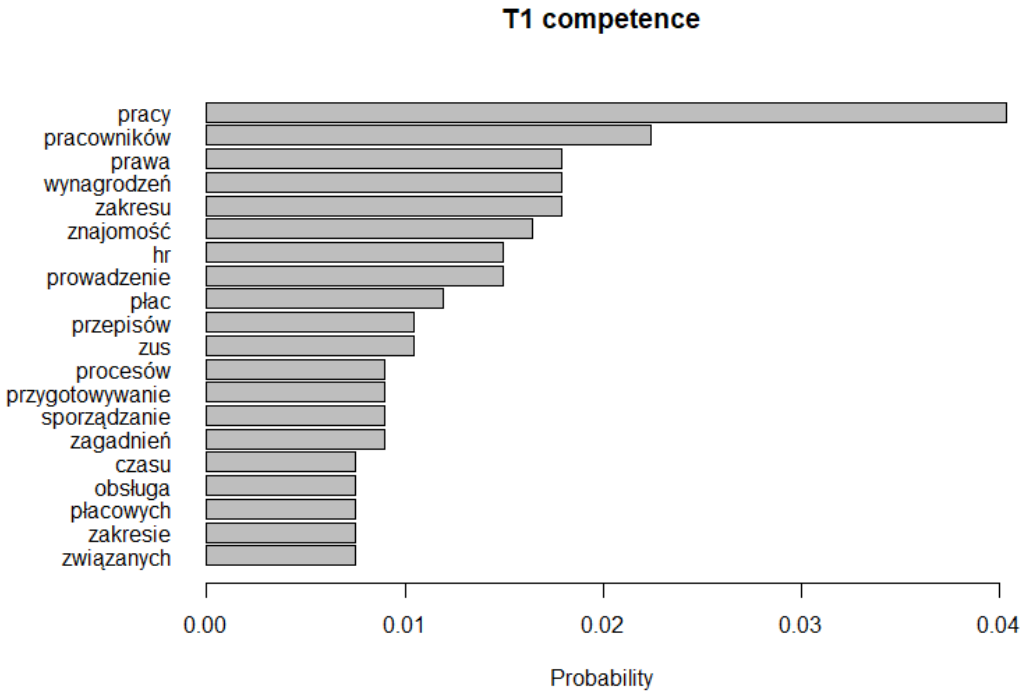

Fig. 2. The distribution of the most important words for the competence T_1.

Next the full set of job offers related to HR area was analysed with the use of the *LDA* algorithm. Calculations were performed with the help of the *topicmodels* package (Grün, Hornik, 2011). During the calculation process 40 topics were identified. Next the topic-competence similarity matrix was estimated with the use of the cosine formula. In the last step of calculation the **K** matrix was generated. To show the results the analysis of an exemplary job offer was performed. The test was prepared in Polish:

"Zadania: Zarządzanie bazą kandydatów. Tworzenie i publikacja ogłoszeń rekrutacyjnych. Selekcja aplikacji. Kontakt z kandydatami. Zarządzanie kalendarzem rekrutacji. Udział w rozmowach rekrutacyjnych. Generowanie raportów rekrutacyjnych. Wymagania: Znajomość języka angielskiego na poziomie pozwalającym bezproblemową komunikację. Zainteresowanie tematyką z zakresu rekrutacji. Bardzo dobre umiejętności interpersonalne. Bardzo dobra organizacja pracy. Mile widziane doświadczenie w pracy w dziale HR. Pracę w młodym, dynamicznym i otwartym na niekonwencjonalne rozwiązania zespole. Pracę w branży nowych technologii. Możliwość rozwoju w ramach wewnętrznych struktur organizacyjnych Integrację w i poza pracą"

The significance of competences in the above offer is presented in the Fig. 3.

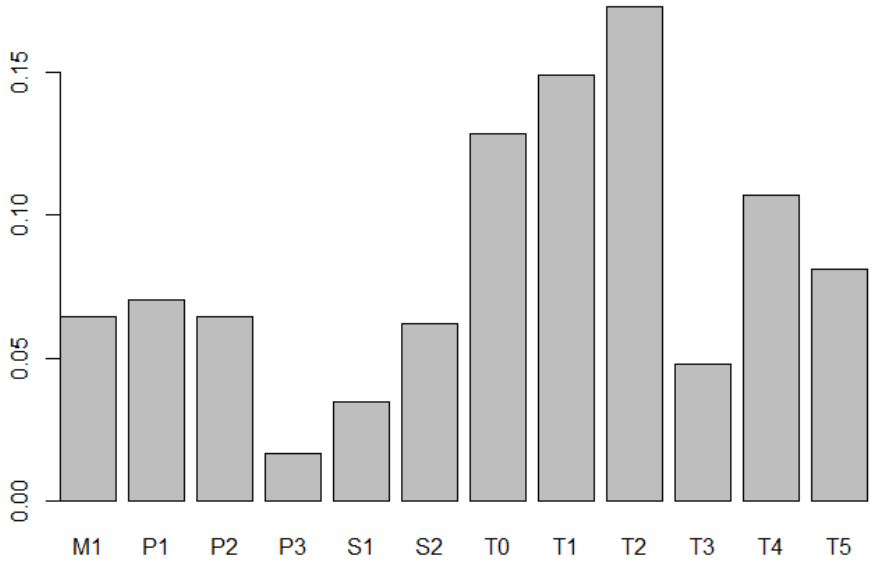

Fig. 3. The significance of competences in the exemplary job offer.

The results shows that competences T_0, T_1 and T_2 were recognized as the most important in an offer.

Conclusions

In the paper the method for job offer mapping from the space of semantic components (topics) identified by the *LDA* algorithm to the space of competences related to the HR area

was proposed. Experience gained from the study allows to point out main advantages of the proposed procedure. Using the method proposed in the paper job offer are presented in the space of competences instead of the space of *LDA* topics. It facilitates the process of automatic interpretation of employers' expectations towards candidates for employment. The proposed model is composed of two sub-models: *LDA* model for topics identification and classification model for mapping documents into the space of competences. These two elements are mutually independent. It means that changes made in one of them do not results in changes of the other. The *LDA* model works as a filter which reduces information noise in the input data by transforming original job offer from the space of words to the space on topics. It seems that the proposition presented in the paper can be helpful for analysis of processes existing on the labour market and can refine education policy of universities and other educational institutions.

Unfortunately, also some disadvantages were noted. Firstly, it should be underline that model creation process may be time-consuming. It is caused by the necessity of manual creation of text files into training set used for classification model building. Secondly, the adjustment of main model parameters (including established taxonomy of competences, the number of topics for the *LDA* model and the formula for similarity calculation between topics and competences) may be demanding and may extent the time necessary for model preparation. All calculations presented in the paper can be performed with the use of R software.

Acknowledgements

The project has been funded by the Cracow University of Economics within the Statutory Fund of the Faculty of Management.

References

Alghamdi, R. & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*. *6*(1), 147-153.

Blei, D., Ng, A. & Jordan, M., (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3), 993–1022.

Feinerer, I., Hornik, K. & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54.

Filipowicz, G. (2014). *Zarządzanie kompetencjami. Perspektywa firmowa i osobista*. Warszawa: Oficyna Wolters Kluwer Business.

Grün, B. & Hornik, K., (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(1), 1–30.

Manning, C. D. & Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA. MIT Press.

Oczkowska, R., Wiśniewska S. & Lula P., (2017). Analysis of the Competence Gap Among Vocational School Graduates in the Area of Smart Specialization in Poland. *International Journal for Quality Research*, *11*(4), 945-966.

Salton, G., Wong, A. & Yang, C., (1975). Vector-Space Model for Automatic Indexing. *Communications of the ACM*. *18*(11), 613-620.

Whiddett, S. & Hollyforde, S. (2003). *A Practical Guide to Competencies. How to Enhance Individual and Organisational Performance*. London: CIPD.

Whiddett, S. & Hollyforde, S. (2007). *Competencies*. London: CIPD.

Wickham, H., (2016), *Package 'rvest'*, https://cran.r-project.org/web/packages/rvest/rvest.pdf

Wright, M., Turner, D. & Horbury, C. (2003). *Competence Assessment for the Hazardous Industries. RR086*. Sudbury: HSE Books. http://www.hse.gov.uk/research/rrpdf/rr086.pdf