SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2020
ORGANIZATION AND MANAGEMENT SERIES NO. 142

THE ROLE OF WORD AND N-GRAM FREQUENCY ANALYSIS IN INFERENCE OF THE CONTENT OF SCIENTIFIC PUBLICATION

Iwona ZDONEK

Silesian University of Technology, Faculty of Organization and Management, Zabrze; Iwona.Zdonek@polsl.pl, ORCID: 0000-0002-3377-0904

Purpose: The paper presents an analysis of a scientific publication with regard to the frequency of words and n-grams. The research problem addressed was the question to what extent the text mining analysis of a scientific publication will allow to infer its content.
Design/methodology/approach: The main research method is the analysis of tokenized text using word count functions, bigrams, and trigrams in selected sections of a scientific publication. The results of text mining analysis were compared with the classic, non-automated text analysis of the publication. The presented study is a pilot project in the form of a case study.
Findings: The proposed method of analyzing a scientific text using an analysis of the frequency of words and n-grams enables inference of the content of the paper with regard to the names of variables involved in the study, the statistical apparatus used and the key literature cited. It should be observed, however, that the discussed method does not make it possible to establish which variables are moderators and which are mediators.
Originality/value: In this paper, the text mining technique was used differently in the discussed study than in previous works. The publication was not examined in its entirety, as previous researchers did, but text mining analysis was applied to individual parts of the paper, i.e. the part discussing theoretical foundations of the research and the part presenting the research method, research results, and their discussion. This allowed for obtaining more precise results regarding the content of the publication.

Keywords: text mining, R in text mining, n-grams, scientific publication analysis.

Category of the paper: A case study.

1. Introduction

With the advent of Web 2.0 and social media, the number of unstructured text data has increased. It is estimated that the number of data saved in text files is 85-90% of all data existing worldwide (Hotho, Nürnberger, Paaß, 2005). To facilitate its analysis, a text mining technique was proposed, which emerged in the 1990s. It is used to analyze texts in order to extract unstructured information contained in data set (Szymańska, 2017). It is, thus, a text mining

technique that automatically extracts previously hidden, unknown and potentially useful information from a large amount of unstructured text data in a scalable and repeatable way (Fan et al., 2006; Frawley et al., 1992). Text mining, therefore, provides information obtained as a result of document analysis and is treated as a special application of statistics (Zwierzchowski, 2017). It is assumed that text mining is interdisciplinary in nature and is related to data mining, natural language processing, information retrieval, statistics, linguistics, mathematics and computer science. The use of text mining is also interdisciplinary. This method is used both in technical, social and natural sciences (see Allahyari et. al., 2017, Berezina et. al., 2016, Fleuren and Alkema 2015, Krallinger et. al., 2017, Boussalis and Coan 2016, Debortoli et. al., 2016, Ngai et al. 2016, Szymańska 2017).

Researchers using text mining in their work emphasize that it was used on large data sets, for simultaneous analysis of many works. However, the usefulness of text mining for a single work is also interesting. Therefore, the research problem undertaken in this paper concerns the answer to the question:

To what extent will text mining analysis of a scientific publication enable inference of its content?

Exploration of text data begins with the general acquisition and standardization of text data. This means obtaining and unifying documents that will be analyzed. Standardization usually involves transformation of the document into a text file, removal of formatting characters, and standardization of the coding characters. The collected documents are then subjected to the stage of tidying the data contained therein. Data tidying, according to Silge and Robinson (Silge, and Robinson, 2019), begins with the tokenization of a text data set. It involves dividing the analyzed text into tokens, i.e. significant units of text that we want to analyze. An example of tokens can be words in the analyzed text. The tokenization process should also include replacement of all uppercase words with lowercase words, removal of the so-called white spaces and punctuation marks. The need to replace all words with lowercase words results from the requirement to standardize their spelling, which is important from the point of view of future analysis. This is done so that when counting the frequency of words in the text at a later stage, the count function does not count the same words like e.g. “project” and “Project” separately. Tokenization is followed by the stage of removing words irrelevant from the point of view of analysis, i.e. stop words. Stop words are words and phrases that give the text meaning and complement it, but do not provide specific information on their own. They include parts of speech such as pronouns, conjunctions, prepositions. Therefore, such words should be removed from the set of analyzed words, as they will not bring relevant content to the analysis. In addition to tokenization and elimination of words irrelevant for analysis, textual data tidying also uses stemming, i.e. finding word cores (e.g. words connect, connected, connecting, connections have a common core “connect”). The processes of tokenization, elimination of stop words and stemming are referred to as preprocessing methods (Vijayarani et al. 2015; Allahyari et al.,

2017). Text analysis begins after the preprocessing methods stage. Word and n-gram frequency and word correlation analyzes are performed the most frequently on tidied text.

2. Methods

Text analysis was performed in R using the tidytext, dplyr and tidyr packages. A publication on human resource management "Supervisor motivating styles and legitimacy: moderation and mediation models" was selected for the pilot study (Kanat-Maymon et al. 2017). It was assumed that the main sections of the document would be analyzed, i.e. theoretical literature background and the methods, results and discussion sections to be analyzed together. Therefore, separate text documents have been created for the mentioned sections of the article. The reason for creating a joint document for the last three sections of the paper was the small volume of some of them and the repeatability of the vocabulary associated with the research method used, especially the statistical apparatus.

After importing data from a text file using the readLines function, the data was recognized as character type. Then it was transformed into a data frame using the tibble command from the tidyr package (see Fig.1).

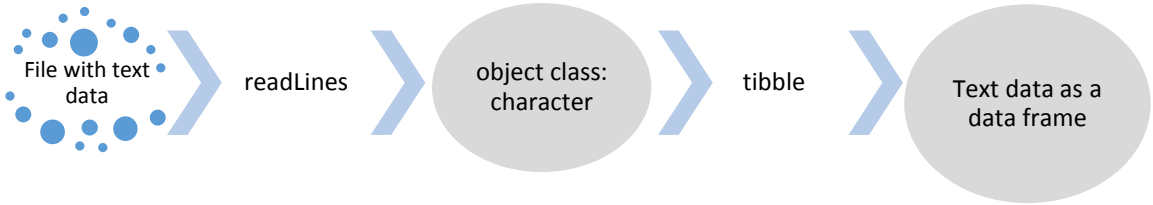

Figure 1. Conversion of text data into data frame. Own study based on: (Silge, and Robinson, 2019).

The created data frame was then tokenized and tidied from words irrelevant for further analysis (stop words). The result was also a data frame (see Figure 2). Tokenization was performed based on words, bigrams and trigrams (see Figure 3), which were then counted and visualized on bar charts. In the process of data tidying, stemming (finding word cores) was not performed as this would not allow the analysis of words and n-grams for parts of speech, which would disturb the analysis.

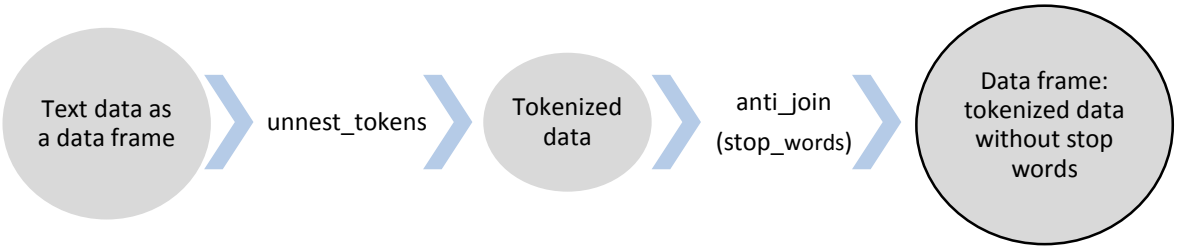

Figure 2. Tokenization and tidying of the data frame of words irrelevant for further analysis. Own study based on: (Silge, and Robinson, 2019).

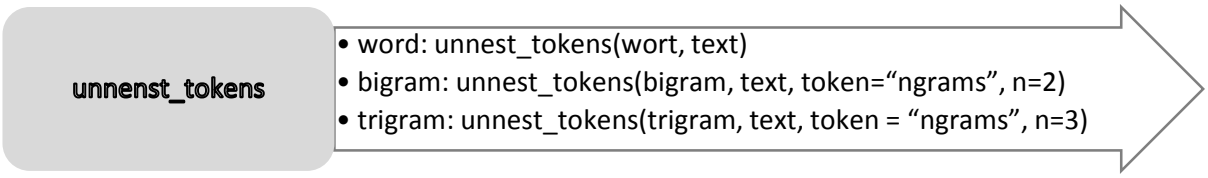

Figure 3. Tokenization of words and n-grams. Own study based on: (Silge, and Robinson, 2019).

The research results obtained with the text mining techniques discussed above were compared with classic text analysis based on text reading. Comparison of the results of both types of analysis enabled conclusions concerning the quality of text mining analysis.

3. Results

The scientific publication analyzed was entitled: "Supervisor motivating styles and legitimacy: moderation and mediation models" (Kanat-Maymon et al. 2017). The keywords presented by the author are: motivation, legitimacy, self-determination. Figures 4-5 present the visualization of the results of text mining analysis for words, bigram and trigrams made for the Theoretical background section, while in Figures 6-7 the visualization for the results of the Methods, Results and Discussion section analysis.

Analysis of the section Theoretical background

Since the title of this paper indicates that a model is developed therein, the variables of this model were selected by searching for nouns in the list of the most common words. Analysis of the word counts presented indicates that the model developed in the paper is associated with the following variables: "legitimacy" (48 occurrences) and "motivation" (23 occurrences), "autonomy" (25), "style" (25), and "employees" (22). Three of these terms also appear in the title (legitimacy, motivating, style) and two (legitimacy, motivating) in the keywords provided by the author. Further analysis of the frequency of words also enabled identification of such variables as: "authority", "supervisor", "outcomes", "support", "depression", "job", "burnout", "engagement", "satisfaction".

Bigram analysis provides further information. The motivational styles discussed in the publication are those related to the terms "autonomous motivation", "autonomy support", "controlling motivating", and "job satisfaction". The bigrams "perceived legitimacy" and "perceived supervisor", "supervisor legitimacy" occurred often (7 times), specifying the variable "legitimacy" and connecting it with the variable "supervisor". Repeated appearance of "positively linked" in bigram suggests that the theoretical background concerned development of the notion of the influence of variables (positive correlation) in the created model.

Trigram analysis further specifies the concepts associated with "legitimacy" by identifying "perceived supervisor legitimacy". Trigrams also show a great deal of focus on development of the topic towards "autonomy supportive motivating". They also indicate "sdt based motivating" (i.e., Self-Determination Theory of Motivation) as the main theory of motivation on which theoretical considerations are based. Bigram and trigram analysis shows that the literature cited in the paper the most frequently Tyler's publications from 1997 and 2005 (8 times each). The papers of Deci from 2017 (7 times) and from 2005 (6 times) were also often cited. It can therefore be assumed that the theoretical research background of the analyzed publication was inspired by these works. In addition, publications from 2000, 2001 and 2013 were often cited, although bigrams did not indicate who was their author. Instead, they showed that the authors: Kanat Maymon (i.e. one of the authors of the analyzed paper) and Ryan made a significant contribution to the theoretical foundation of the analyzed publication.

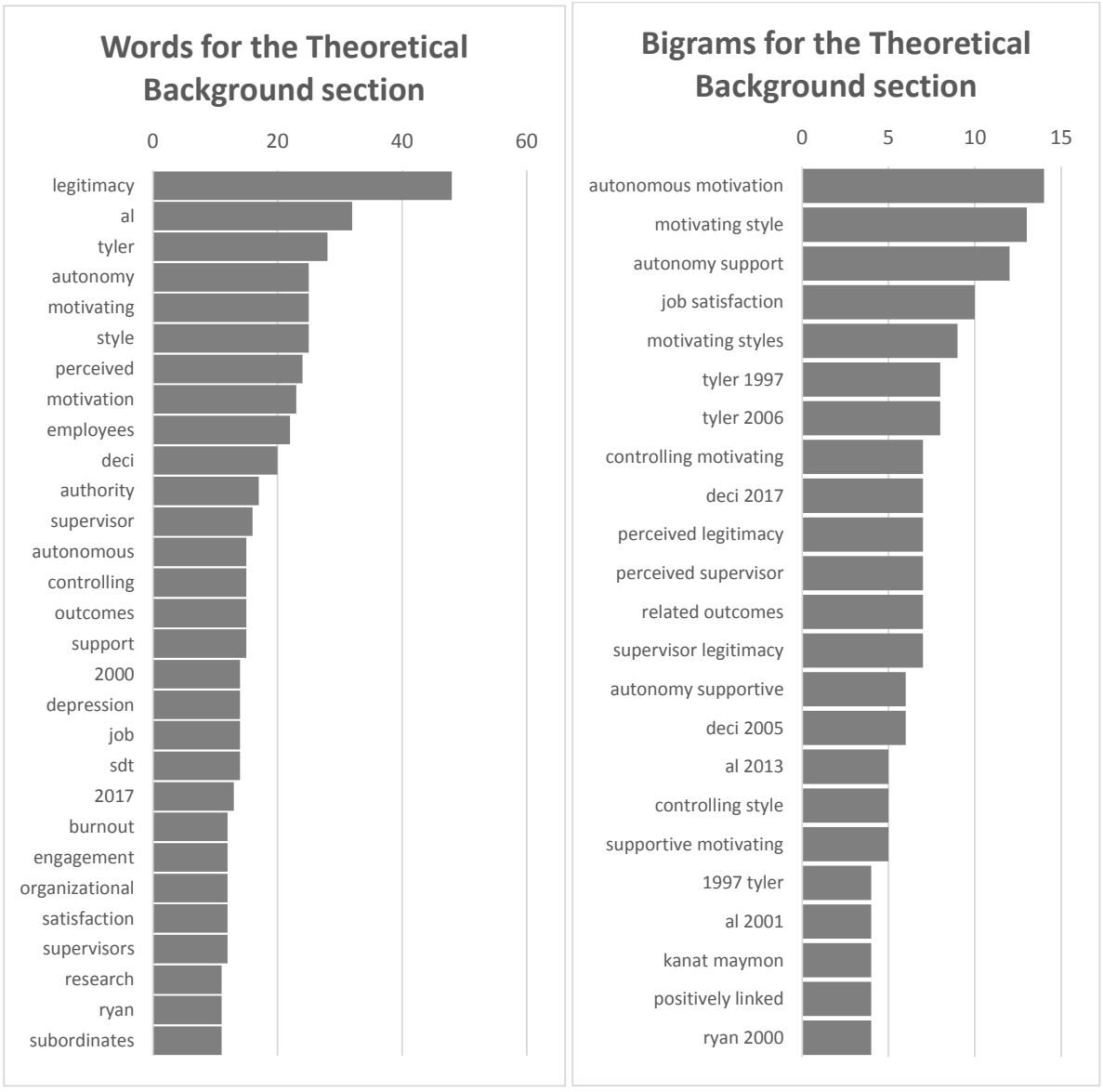

Figure 4. Words and bigrams for the section Theoretical background. Own work.

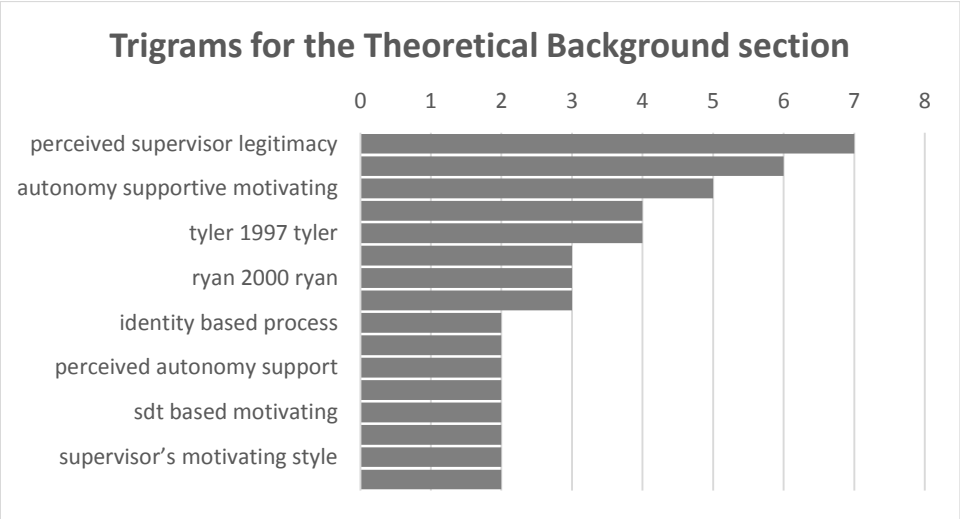

Figure 5. Words and bigrams for the section Theoretical background. Own work.

Classic text analysis (i.e. by reading) has shown that the variables in the model under development are: "supervisor's autonomy motivating style", "supervisor's controlling motivating style", "job satisfaction", "affective commitment", "engagement", "burnout", "depression". Therefore, comparing the results obtained from the list of word and n-gram frequency, it can be observed that the text mining analysis identified all model variables except the "affective commitment" variable. Therefore, it can be said that the presented method of searching for model variables based on the highest frequency of words and n-grams is highly effective.

Analysis of the section Methods, Results and Discussion

Analysis of the frequency of words, bigrams and trigrams in the section Methods, Results and Discussion shows that the variable associated with all variables was "legitimacy" (this is indicated by its high frequency of occurrences, n=55 and confirms the presence of this variable in the title of the publication). Bigrams indicate that "legitimacy" is more precisely referred to as "perceived legitimacy". Variables that are associated with it are variables related to the terms:

- "autonomy support", "autonomous motivation", "autonomy supportive", "autonomy supportive style", "supervisor autonomy support", "supervisor motivating style", "perceived autonomy support",
- "controlling style", "controlling motivation", "supervisor controlling", "controlling motivating style", "supervisor controlling style",
- "supervisor motivating style",
- "job satisfaction", "satisfaction engagement",
- "employee outcomes".

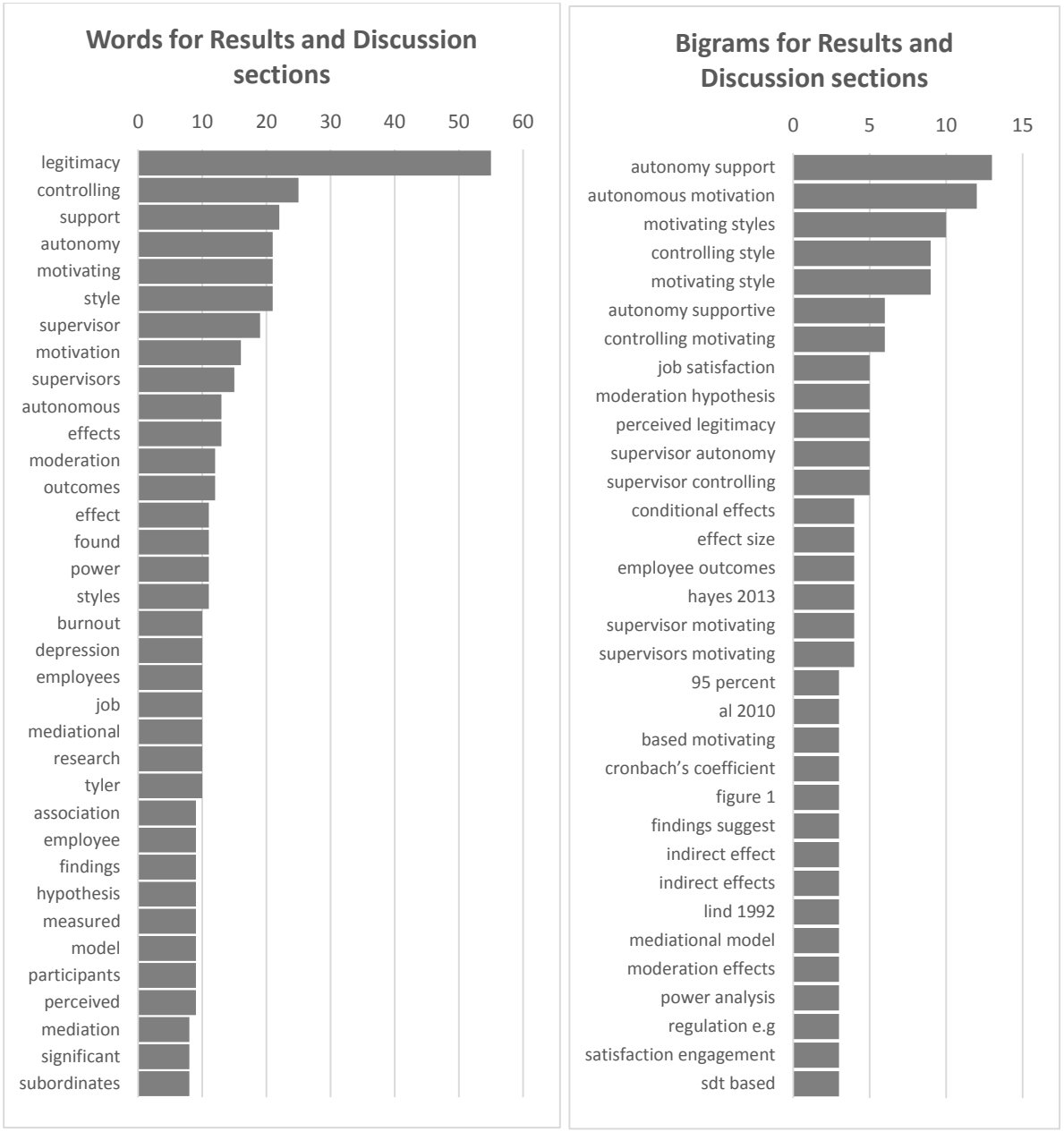

Figure 6. Words and bigrams for the section Methods, Results and Discussion. Own work.

The words and bigrams also feature terms related to moderating and mediating variables ("moderation", "mediational", "mediation", "moderation hypothesis" "mediational model", "moderation effects", "indirect effect") and the terms used regarding statistical inference ("significant", "95 per cent", "cronbach's coefficient", "significant positive association"). They indicate the statistical methods used. As regards the literature cited in the Methods, Results and Discussion section, the papers referred to the most frequently are: Hayes, 2013, Lind, 1992 and Tyler, 1997.

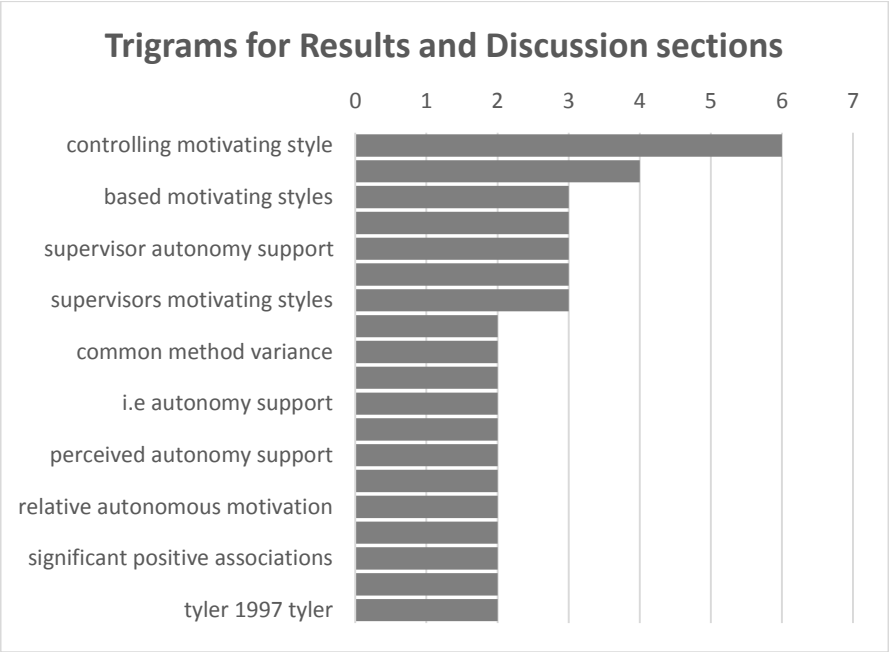

Figure 7. Trigrams for the section Methods, Results and Discussion. Own work.

Classic text analysis by reading confirmed the variables indicated by text mining, the statistical methods used and the focus of the research on the search for mediators and moderators. However, comparing the results of the described text mining analysis with the classic text analysis, one can notice the advantage of the classic method, which gives results not only in the form of variables, but also describes the positive or negative impact of individual variables on each other (specifically on the "perceived legitimacy" variable and specifies which variables are mediators or moderators. The presented text mining analysis based on the frequency of words and n-grams did not enable drawing such conclusions.

4. Discussion

The text mining technique was used differently in the discussed study than in previous works (see Szymańska, 2017; Wyskwarski, 2017). The publication was not examined in its entirety, as previous researchers did, but text mining analysis was applied to individual parts of the paper, i.e. the part discussing theoretical foundations of the research and the part presenting the research method, research results and their discussion. Three sections of the publication were analyzed together, due to the fact that the vocabulary associated with the statistical apparatus applied was repeated in each of them. Moreover, some of these sections were not very extensive, and most text mining researchers suggest working with large text sets (see Allahyari et al., 2017). The stemming process, recommended by some researchers, was not used (see Allahyari, and Kochut, 2015, Allahyari et al., 2017, Vijayarani, 2015) as its implementation would make it impossible to recognize the parts of speech returned by the word

count and n-gram functions used. And the analysis of parts of speech was crucial for concluding on model variables. Since the title of the publication suggested that it relates to the construction of a model, the approach based on the analysis of parts of speech enabled identification of variables of the created model in the analyzed publication. Techniques proposed by Silge and Robinson (Silge, and Robinson, 2019) based on counting word frequency and n-gram frequency gave very good results in this study. Unfortunately, these techniques did not work in the analysis of the combined text of sections: Methods, Results and Discussion. They only allowed to infer the statistical apparatus used, the names of variables of the model, and the literature cited in this part of publication. They did not enable inference on how variables affect each other. Perhaps the postulate on working on large text sets in this case proved to be unfavorable and approaching the said sections separately would give better results. It should be examined in subsequent studies.

5. Conclusion

Due to the fact that the title of the analyzed paper suggested the construction of a model, it was essential to answer the question what variables make up this model. A significant role in finding these variables was played by word frequency analysis, which was the starting point for inference of the model variables developed in the paper. Bigram and trigram analyzes deepened inference and presented the variables in a more precise way. N-grams (bigrams in particular) also revealed which of the cited publications had a significant contribution to the theoretical foundation of the model presented in the paper. The analysis of the frequency of words and trigrams supported the inference from bigrams. The analysis revealed the following variables:

- "autonomy support", "autonomous motivation", "autonomy supportive", "autonomy supportive style", "supervisor autonomy support", "supervisor motivating style", "perceived autonomy support",
- "controlling style", "controlling motivation", "supervisor controlling", "controlling motivating style", "supervisor controlling style",
- "supervisor motivating style",
- "job satisfaction", "satisfaction engagement",
- "employee outcomes".

In turn, the analysis of some words and n-grams from the Methods, Results and Discussion section only confirmed the names of the variables studied and the statistical methods used. It was known, for example, that the confidence level was 0.95, the Cronbach coefficient was calculated, and the most common words and bigrams presented pointed to the search for mediating and moderating variables (mediators and moderators). However, on the basis of the

performed text mining analysis, it could not be concluded whether the influence of individual variables on each other was significant.

The conclusions about the works cited in the publication were an interesting discovery. An analysis of the frequency of words and n-grams enabled the identification of publications that had a significant impact on the analyzed paper, both in the part devoted to theoretical foundations of the model and the part concerning research results (mainly discussion of these results). Bigram and trigram analysis showed that the literature cited in the paper the most frequently Tyler's publications from 1997 and 2005. The papers of Deci from 2017 and from 2005 were also often cited. It can therefore be assumed that the theoretical research background of the analyzed publication was inspired by these works.

It can, therefore, be concluded that the proposed method of analyzing a scientific text using an analysis of the frequency of words and n-grams enables inference of the content of the paper with regard to the names of variables involved in the study, the statistical apparatus used and the key literature cited. The time needed for analysis with this method included: creating files containing the part devoted to theoretical background and the part devoted to results and their discussions, as well as time devoted to counting the frequency of words and the frequency of n-grams. Creating files was done by copying the relevant parts of the paper (less than 10 minutes), while counting the frequency of words and n-grams took seconds. This is a significantly shorter time than the time that would have to be spent to read or at least skim over the publication. It should be observed, however, that the discussed method does not make it possible to establish which variables are moderators and which are mediators (in contrast to the method based on classic text reading). Therefore, further development of text mining techniques should be directed towards enabling inference in this regard.

References

1. Allahyari, M., and Kochut, K. (2015). *Automatic Topic Labeling using Ontology-based Topic Models.* Available online https://www.researchgate.net/publication/300408939_ Automatic_Topic_Labeling_Using_Ontology-Based_Topic_Models.
2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D, Gutierrez, J.B, and Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.* Available online https://arxiv.org/pdf/1707.02919v2.pdf.
3. Berezina, K., Biligihan, A., Cobanoglu, C., and Okumus, F. (2016). Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. *Journal of Hospitality Marketing & Management, 25(1)*. doi: https://doi.org/10.1080/ 19368623.2015.983631.

4. Boussalis, C., and Coan, T.G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*, pp. 89-100. doi: https://doi.org/10.1016/j.gloenvcha. 2015.12.001.
5. Debortoli, S., Müller, O., Junglas, I., and vom Brocke, J. (2016). Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems, 39,* doi: https://doi.org/ 10.17705/1CAIS.03907.
6. Fan, W., Wallace, L., Rich, S., and Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49(9)*, pp. 76-82. Available online https://www.research gate.net/publication/220421836_Tapping_the_Power_of_Text_Mining.
7. Fleuren, W., and Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods, 74,* pp. 97-106. doi: https://doi.org/10.1016/j.ymeth.2015.01.015.
8. Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1992). Knowledge discovery in databases: An overview. *AI Magazine, 13(3)*, pp. 57-70.
9. Hotho, A., Nürnberger, A., and Paaß, G. (2005). A Brief Survey of Text Mining. Available online https://www.researchgate.net/publication/215514577_A_Brief_Survey_of_Text_ Mining.
10. Kanat-Maymon, Y., Mor, Y., Gottlieb, E., and Anat Shoshani, A. (2017). Supervisor motivating styles and legitimacy: moderation and mediation models. *Journal of Managerial Psychology*. doi: https://doi.org/10.1108/JMP-01-2017-0043.
11. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., and Valencia, A. (2017). Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Review, 117, 12,* pp. 7673-7761, doi: https://doi.org/10.1021/acs.chemrev.6b00851.
12. Ngai, E.W.T., and Lee, P.T.Y. (2016). *A review of the literature on applications of text mining in policy making*. PACIS 2016 Proceedings. 343. Available online http://aisel.aisnet.org/pacis2016/343.
13. Silge, J., and Robinson, D. (2019). *Text Mining with R*. Available online https://www.tidytextmining.com/index.html.
14. Szymańska, A. (2017). Wykorzystanie algorytmów text mining do analizy danych tekstowych w psychologii. *Socjolingwistyka, XXXI*, pp. 99-116. Available online http://dx.doi.org./10.17651/SOCJOLING.31.6.
15. Vijayarani, S., Ilamathi, J., Nithya, and Phil, M. (2015). Preprocessing Techniques for Text Mining – An Overview. *International Journal of Computer Science & Communi-cation Networks, 5(1)*, pp. 7-16. Available online https://pdfs.semanticscholar.org/1fa1/ 1c4de09b86a05062127c68a7662e3ba53251.pdf.
16. Wyskwarski, M. (2017). Text mining w analizie zbiorów publikacji naukowych. *Zeszyty Naukowe Politechniki Śląskiej, seria Organizacja i Zarządzanie, 114*. Avaliable online http://dx.doi.org/10.29119/1641-3466.2017.114.49.
17. Zwierzchowski, D. (2017). *Text mining i narzędzia eksploracji tekstu*. Available online https://www.researchgate.net/publication/313772976_Text_mining_i_narzedzia_eksplora cji_tekstu.