Clustering of Business Organisations based on Textual Data - An LDA Topic Modeling Approach

Ferenc Tolner[1,3], Márta Takács[3], György Eigner[2] and Balázs Barta[1]

[1]*Pannon Business Network Association*, Szombathely

[2]*University Research and Innovation Center, Physiological Controls Research Center, Óbuda University*

[3]*Óbuda University Doctoral School of Applied Informatics and Applied Mathematics*, Budapest, Hungary

Email: ferenc.tolner@am-lab.hu, takacs.marta@nik.uni-obuda.hu, eigner.gyorgy@nik.uni-obuda.hu, balazs.barta@pbn.hu

2021 IEEE 21st International Symposium on Computational Intelligence and Informatics (CINTI) | 978-1-6654-2684-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CINTI53070.2021.9668337

***Abstract*—Textual data provides a new perspective and a huge potential with additional information in analysing and segmenting business organisations. Statistical "hard data" is often too general or even misleading and might be affected by several exogenous and endogenous factors while questionnaire or survey related "soft data" is hardly available or can be biased by the interviewees position in the organisation or by its own personal orientation. On the other hand, besides the aforementioned information sources business organisations, education- and research institutions etc. provide many times textual data on themselves as well, that can further contribute to the understanding of the investigated population. In this paper a topic modeling of 51 Central European business-, educational- and research organisation has been performed by Latent Dirichlet Allocation (LDA). The investigated organisations were partakers of an online survey where their textual organisational descriptions were collected together with basic geographical and industry related data. Based on the result a grouping of the stakeholders has been implemented and an LDA based methodology has been tested in order to further support cluster-forming efforts of business- and other type of organisations within the Central European region.**

***Index Terms*—SME, resilience, cross-border relationships, business clusters, LDA, Latent Dirichlet Allocation, topic modeling**

I. Introduction

The aim of our present work is to elaborate a methodology applied on freely or relatively easily obtainable organizational data in order to contribute to more efficient international cross-border partnership establishments in the Central European region. By achieving such long-term, successful co-operations of various organizations the overall competitiveness of the Central European region is expected to increase as well the resilience against economic turbulences [1].

With the huge amount of available textual inputs a more detailed insight into data can be gained that can serve with automatised and aggregated feedback on customer satisfaction, new trends or can reveal hidden, unstructured semantic information embedded in texts. Topic modeling belongs to the unsupervised Artificial Intelligence learning methods and has diverse application possibilities in multidisciplinary fields of natural- and social sciences. Text summarising, classification, sentiment analysis, prediction by topic popularity investigation in time and topic clustering are all belonging to the various application areas of topic modeling [2], [3], [4],, [5].

In economics the description of financial systems, efficiency of companies etc. is often described by indicator numbers such as liquidity ratio, annual revenue etc., however these data are commonly not freely obtainable. Consequently other data sources has to be used, as textual data, that can provide additional information on the direction of the development processes. Since large amount of textual data is hard to comprehend visualisation techniques like *Word Clouds* are required for the understanding what the documents are about [6], [7].

Latent Dirichlet Allocation (LDA) is a useful text mining technique that enables an insight into large amount of unstructured texts. It assumes the documents to be a discrete probability distribution of topics and similarly the topics a discrete probability distribution of words. The documents and words are the observable subjects, while the topics remain latent that has to be unveiled. Among the hardships of LDA it has to be mentioned that the discrete distribution of words in the topics and documents is often pretty much skewed. Therefore subjective specification and removal of "stop words" that contain less topical content is essential and a high sensitivity of hyper-parameters and topic numbers is expected [8], [9].

LDA was first published in 2003 as a probabilistic model to reveal hidden semantic structures in textual data. For this purpose the creation of a document-term matrix is needed which encodes the occurrence frequencies of the words in different documents in a normalized form. Since there might be millions of different words in a larger document the size of the document-term matrix can became tremendous and especially sparse and techniques for text pre-processing are desirable to reduce variability of words while keeping the content [3]. Further description on the applied pre-processing techniques will be outlined in Section II. and the overall topic modeling process can be seen on Fig. 1.

Topic modelling techniques such as LDA offer an algorithmic and automatized solution to organize, annotate and manage large amount of unstructured documents that would be unimaginable by hand. However it is still an open question and heavily researched how topic models can be compared based

The publication of this article has been supported by the Robotics Special College via the "NTP-SZKOLL-20-0043 Talent management and community building in the Robotics Special College of Óbuda University" project and by the Pannon Business Network Association.

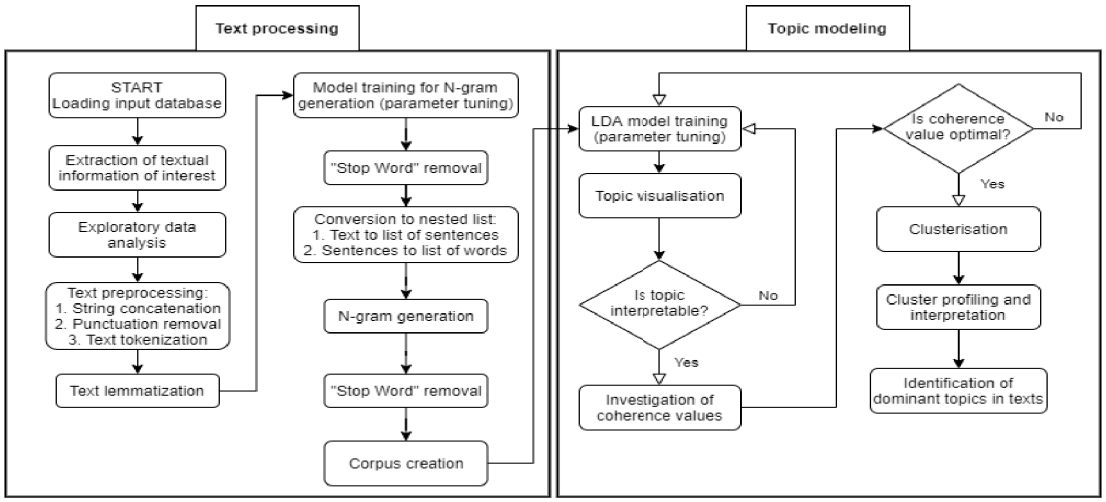

Fig. 1: Flowchart of the applied topic modeling procedure.

on how interpretable they are, how they fit the user's goals, which is addressed as *model checking problem*. Therefore the question still holds which model can be the most appropriate for given text samples. Thus the results shall be treated accordingly [10].

II. The Collected Data

The data collection was realized by an online survey that had the aim to bring together different organisations and gather information on their interests and special focuses. By encouraging international actors to engage in cross-border collaborations or form business clusters information dissemination and risk reduction can be achieved among the members, may they be from competitive or complementary business fields [11], [12], [13]. Such clustering initiatives are expected to increase the competitiveness of the individual actors and of the Central European region as well, therefore highly supported by the European Council [14].

Besides textual information –that forms the main focus of the present study– general data on business branches with NACE[1] numbers, addresses and feedback on a multiple selection choice survey were collected from 51 Central European business, educational- and research organisations, that intended to map professionalindustrial areas on which the participants were willing to broaden their present limits of knowledge and were ready to participate in an international cross-boarder collaboration.

For the exploratory data analysis the Anaconda 3 framework was utilised that is a freely available software environment on Windows, Linux and Mac OS X to simplify scientific computing tasks [15]. The main attributes of the gathered general information can be seen on Fig. 2. and on Fig. 3.

The geocoding[3] and visualisation of address information of the attendees were performed by the open source *geopandas 0.8.0* python toolbox [16] (see Fig. 4.). As the geographic localisation shows, partakers from 7 Central European countries and from 18 NUTS2[4] regions were involved in our study with a geographically mixed distribution regarding organisational types.

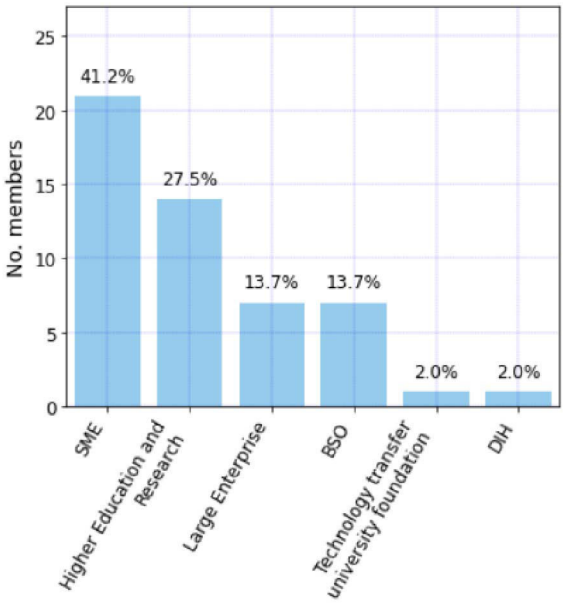

Fig. 2: The distribution of the organisational types[2]of the 51 Central European participants.

The textual information on the participants involved:

- General description of the participating organisation and their main activities.
- An account on their competences and field of expertise

[1]Standard classification of economic activities applied in the European Union.

[2]BSO: Business Support Organisation, DIH: Digital Innovation Hub

[3]The process of generating latitude and longitude coordinates and thereby exact geographic locations from addresses on a map.

[4]Nomenclature of Territorial Units for Statistics: Geocoding standard for referencing the subdivisions of EU member states for statistical purposes.

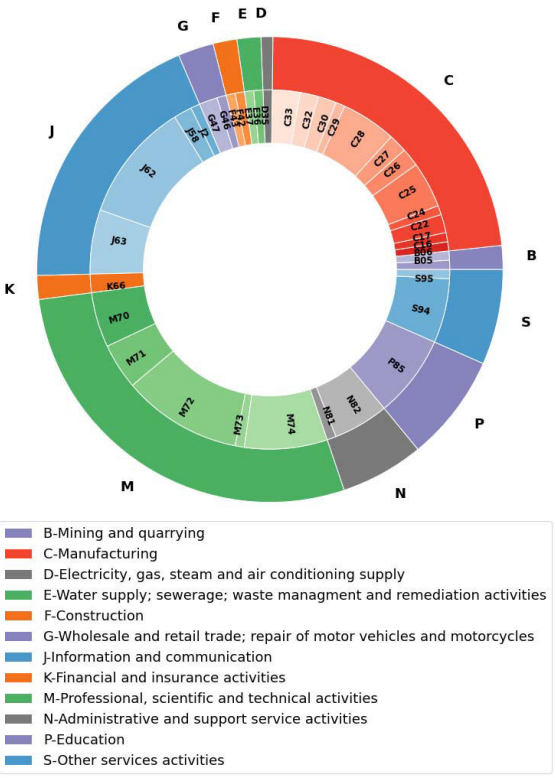

Fig. 3: The distribution of the different pursued economic activities at present by the 51 Central European participants.

that included their products and services which formed the basis of their value-added benefit.

- Brief summary on the strengths and outcomes that might entitle the partakers to be involved in further co-operation projects.
- Main challenges faced in their region.
- Description of innovative solutions that contributed to overcome the challenges faced in their region.
- Key value-added benefit sought by taking part in an international cross-border co-operation activity.

A summary boxplot on the number of words in each text provided for our topic modeling purposes is outlined on Fig. 5. This indicates mostly texts in length of 100 words are available apart from a few exceptions where longer descriptions were given.

III. Results, Discussion

For the exploratory data analysis and text processing of the provided textual information the open source *python NLTK 3.6.2.* toolbox for Natural Language Processing tasks and the *python wordcloud 1.8.1.* toolbox for text visualisation have been utilised [17], [18]. Further text processing has been outlined by using the *spaCy 3.0* and *gensim 4.0.1* python modules [19], [20]. Besides text processing the *gensim 4.0.1* module has been applied for the topic modeling purposes as well. Additionally to the aforementioned software packages for further visualisation and interpretation of the results we

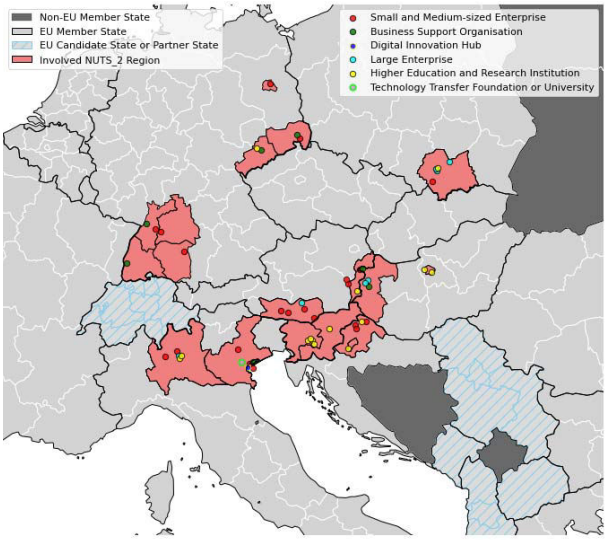

Fig. 4: The geocoded geographical distribution of the 51 Central European participants labeled with the corresponding organisational types and NUTS2 regions concerned.

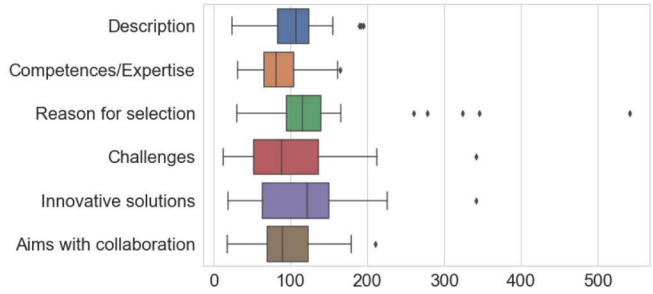

Fig. 5: Exploratory data analysis of the available textual data.

relied on the capabilities of the *pyLDAvis 3.3.1.* module [21], [22].

Since our exploratory data analysis showed that the available text snippets are relatively short and due to the nature of the described topics have overlapping, synonymous and redundant elements per organisation, the strings have been concatenated to a common text in case of every interviewee. Therefore 51 texts arose in our corpus for further text processing and topic modeling. A detailed flow chart of the applied procedure outlined in this section for extracting structured information from the textual inputs is given in Fig. 1.

A representation of the content and main focuses of the investigated texts at significant text processing steps can be given by a "Word Cloud" visualisation (see Fig. 6. and 7.). This serves with a brief summary that the organisations concerned were especially interested in production and development related topics besides technology- and solution oriented matters.

[5]Splitting of text into sentences and sentences to words.

[6]Reduction of words to their corresponding vocabulary form. Removal of pre- and suffixes.

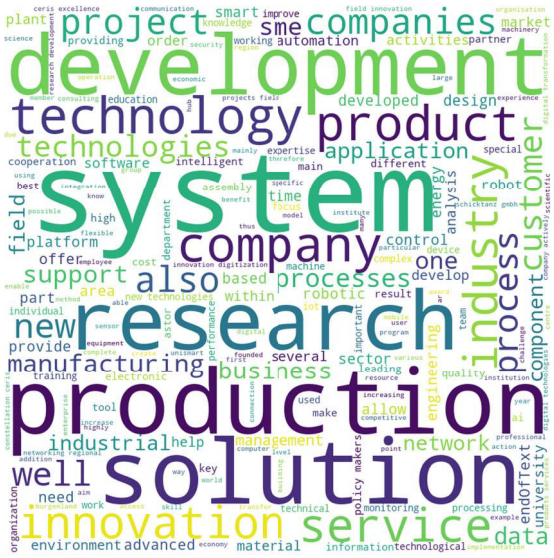

Fig. 6: Word cloud representation of the resulted 51 texts after converting characters to lowercase, removing stop words and punctuation characters performing tokenization[5].

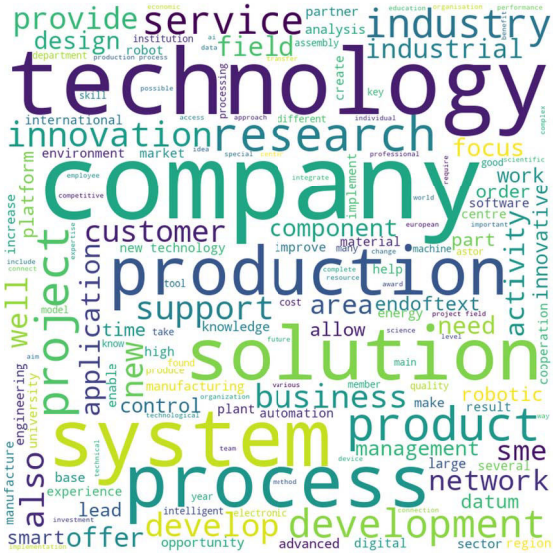

Fig. 7: Word cloud representation of the resulted 51 texts after stemming and lemmatization[6].

After accomplishing the substantial text processing steps in order to reduce the dimensionality of our data and increase computing performance bi- and tri-grams [7] were generated that further increased the interpretability of the text by adding new compound words with modified meaning to the list of already lemmatized words. Such bi- and tri-grams in our investigation among others were: *smart_material*, *machine_learning*, *predictive_maintenance*, *developing_innovative_product* etc. Nevertheless the steps of tokenization, lemmatization and N-gram[8] generation did not produce perfect words this did not cause any trouble to the investigation. At this stage the main goal is dimensionality reduction and the final interpretation has to be done by the researcher staff for whom the meaning of faulty fragmented words in a summarizing visualisation is obvious.

During the LDA topic modeling the number of topics has to be pre-specified. Therefore, in an iterative procedure a parameter sweep has been performed and the topic coherence values have been investigated for the estimation of the optimal topic number. The topic coherence value serves with a simple topic score that gives the level of semantic similarity of high-scoring words in the different texts. This evaluation metric can be utilised for LDA model selection with a given hyper-parameter set on a quantitative basis [2] [23]. Based on the selected C_v topic coherence values the further examination of 8 topics has been performed (see Fig. 8.).

The resulted topics were separately viewed on an inter-topic distance plot offered by the *pyLDAvis 3.3.1.* package where the relation of topics to each other and possible higher level structure of groups of topics can be visualised and analysed. On the left of Fig. 9. a multidimensional scaling of the 8 topics are presented. The size of each bubble corresponds to the percentage in corpus that is about the given topic. On the right the horizontal blue bars represent the overall frequency of the listed words in corpus, while the red bars corresponds to the estimated frequency of each term[9] within the selected topic from the left. The larger the proportion of the red bar compared to the blue indicates a more common usage of the terms within the selected topic.

Having selected the optimum number of topics the dominant topic for each partaker organisation can be determined. In our case this serves with an obvious possibility for the segmentation of the 51 stakeholders into 8 different groups. The number of members in each group is summarized in Table I., while the most relevant keywords for each topic at $\lambda = 0.6$ term relevance factor that is suggested in [21] are listed in Table II.

Cluster ID	No. of members	No. members/Total (%)
Cluster0	16	31.4%
Cluster1	12	23.5%
Cluster2	11	21.5%
Cluster3	5	9.8%
Cluster4	3	5,9%
Cluster5	2	3,9%
Cluster6	1	2%
Cluster7	1	2%
Total:	51	100%

Table I: Number of members within the resulted clusters and the relative cluster sizes compared to the sample size.

The selection of the most relevant keywords are somewhat subjective, since a human decision is necessary in order to

[7]Two- or three word long compound words formed from words that occur often together in text. They are counted as one word in the document-term matrix in order to create more meaningful terms for topic models in case of NLP problems.

[8]Automatic generation of compound words out of N words that occur often together in text.

[9]Nomenclature for pre-processed words.

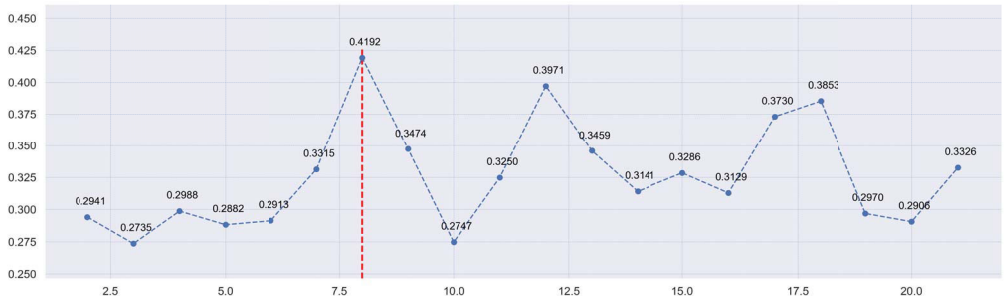

Fig. 8: Determination of the ideal number of topics based on the investigation of coherence values.

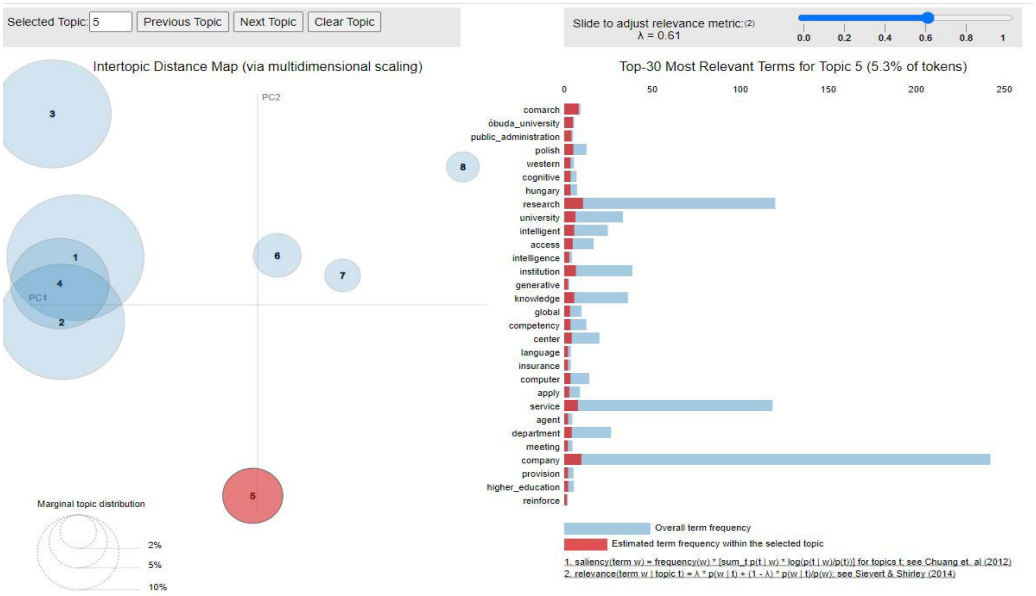

Fig. 9: Visual representation of the resulted topics by the pyLDAvis package.

Cluster ID	Dominant keywords	Ratio of tokens (%)
Cluster0	**technology**, production, product	28.4%
Cluster1	**service**, system, development	23.9%
Cluster2	**customer**, company, solution	20.9%
Cluster3	**application**, robotic, engineering	14.7%
Cluster4	**intelligence**, university, competency	5.3%
Cluster5	**venture**, streaming, talent	3.4%
Cluster6	**furniture**, cluster, wood	1.8%
Cluster7	**engine**, worker, AR	1.6%

Table II: The most relevant keywords for each topic at $\lambda = 0.6$ term relevance factor.

interpret a topic. Therefore three keywords have been selected and the most descriptive and general term marked with bold in Table II. in order to have a clear view on the resulted topics.

Approximately 88% of the tokens of the texts could be assigned to 4 topics, while the remaining 4 topics formed separated smaller groups that can be regarded as ”outlious”.

The distribution of organizational types, NACE categories and locating countries within the clusters given by the topic modeling have been visualised on heatmaps and further interpreted (for the distribution of organisational types see Fig.10.).

Based on the heatmap investigations no correlation were found among any aforementioned variable and cluster indexes, however there was no cause to assume such either. Therefore the used technique served with a different clustering of the partakers as would have resulted not taking into consideration textual information and thus could contribute with further input to the examination of the sample.

The presented method served a thorough and comprehensive overview on the textual information available in an easily interpretable way. This can facilitate non-professionals and decision makers as well in understanding and circumscribing focuses and interests of business organization and support policy makers in more efficient international cross-border cluster generation.

IV. Conclusion and Future Work

In our present work the exploration and topic modeling of textual data provided from 51 Central European business-,

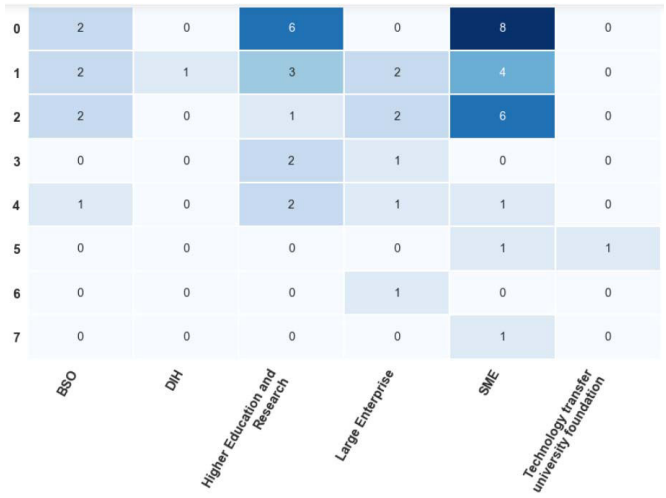

Fig. 10: Distribution of organisational types within the resulted clusters.

education- and research organisations has been outlined. The data concerned were collected via an online survey where basic organisational information and preference options on a possible future co-operation and further education were asked for. The provided textual data incorporated short descriptions on the organisations' focuses, main operation areas, activities, challenges and successes.

The given data served as a basis for the implementation of a topic modeling solution based on *Latent Dirichlet Allocation* (LDA) that can yield an elaborated method for decision makers and policy makers as well in extracting and understanding relevant information from freely available short descriptions of market actors. Due to the increased online activities of the present era the textual information provided and shared has an ever increasing role. Moreover textual information –opposed to numeric data – often easy or even completely free to access that still can encompass valuable content. This further suggests that increased attention on text mining activities shall be paid in the future.

The performed exploratory data analysis and topic modeling served with an overview about the options how textual data can be utilized. Dominant topics and keywords have been specified and a clustering of the 51 stakeholders has been proposed. This approach can contribute to a sophisticated grouping of the members based on their non-trivial interests and descriptions and can lead to a more efficient collaboration throughout future partnership activities.

In a future work our intention is to extend our present findings with the analysis of multiple selection choices collected from the same business organisations on their preferences aforementioned and combine the topic modeling results with a partitioning of the same partakers based on the gathered categorical data. Via elucidating the grouping of companies from different possible aspects realizable in practice an enhanced support of cross-border, international partnership establishment of market actors is feasible.

V. Acknowledgement

This study is a collaborative work of the Pannon Business Network Association and the Óbuda University Budapest. We would like to express our thanks to Mr. Martin Dan and Mr. Géza Éder for their invaluable support throughout of the outlined research.

References

[1] A. Reveiu and M. Dardala, "The role of universities in innovative regional clusters. empirical evidence from romania," *Procedia - Social and Behavioral Sciences*, vol. 93, pp. 555–559, 2013.
[2] S. Rani and M. Kumar, "Topic modeling and its applications in materials science and engineering," *Materials Today: Proceedings*, vol. 45, no. 6, pp. 5591–5596, 2021.
[3] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet Allocation (LDA) for topic modeling of the CFPB consumer complaints," *Expert Systems with Applications*, vol. 127, no. 1, pp. 256–271, 2019.
[4] T. Hayashi and H. Fujita, "Word embeddings-based sentence-level sentiment analysis considering word importance," *Acta Polytechnica Hungarica*, vol. 16, no. 7, 2019.
[5] P. Gnip, L. Vokorokos, and P. Drotár, "Selective oversampling approach for strongly imbalanced data," *PeerJ Computer Science*, vol. 7, p. e604, 2021.
[6] G. Li, X. Zhu, J. Wang, D. Wu, and J. Li, "Using lda model to quantify and visualize textual financial stability report," *Procedia Computer Science*, vol. 122, pp. 370–376, 2017.
[7] B. Madoš and N. Ádám, "Reading volume datasets from storage–using segmentation metadata, for an enhanced user experience," *Acta Polytechnica Hungarica*, vol. 18, no. 5, pp. 187–205, 2021.
[8] H. M.Wallach, D. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," *Advances in Neural Information Processing Systems*, vol. 23, pp. 1973–1981, 2009.
[9] M. Smatana and P. Butka, "TopicAE: A topic modeling autoencoder," *Acta Polytechnica Hungarica*, vol. 16, no. 4, 2019.
[10] G. Beriha , B. Patnaik, S. Mahapatra, and S. Padhee, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.
[11] Y. Kajikawa, J. Mori, and I. Sakata, "Identifying and bridging networks in regional clusters," *Technological Forecasting and Social Change*, vol. 79, no. 2, pp. 252–262, 2012.
[12] Y. Kajikawa, Y. Takeda, I. Sakata, and K. Matsushima, "Multiscale analysis of interfirm networks in regional clusters," *Technovation*, vol. 30, no. 3, pp. 168–180, 2010.
[13] C. Felzensztein, E. Gimmon, and K. R. Deans, "Assessment of safety performance in indian industries using fuzzy approach," *Industrial Marketing Management*, vol. 69, pp. 116–124, 2018.
[14] A. Némethné Gál, "A kis- és középvállalatok versenyképessége - PhD thesis," 2009, István Széchenyi University, Doctoral School of Regional- and Business Administration, Győr.
[15] Computer software. Vers. 2-2.4.0. Anaconda, "Anaconda software distribution," *Web.: https://anaconda.com*, Nov. 2016.
[16] K. Jordahl, "Geopandas: Python tools for geographic data," *Web.: https://github. com/geopandas/geopandas*, 2014.
[17] Bird, Steven, Edward Loper and Ewan Klein, "Natural language processing with python." *O'Reilly Media Inc.*, 2009.
[18] A. Mueller, "Python wordcloud documentation," *Web.: https://amueller.github.io/word_cloud/*, 2020.
[19] "spacy python module," https://spacy.io/usage/linguistic-features, accessed: 2021-06-02.
[20] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.
[21] C. Sievert and K. E. Shirley, "Ldavis: A method for visualizing and interpreting topics," *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70, Jun. 2014.
[22] "pyldavis python module," https://github.com/bmabey/pyLDAvis, accessed: 2021-06-02.
[23] V. Gangadharan and D. Gupta, "Recognizing named entities in agriculture documents using LDA based topic modelling techniques," *Procedia Computer Science*, vol. 171, pp. 1337–1345, 2020.