Topic Modelling in Social Sciences - Case Study of Web of Science*

Maja Buhin Pandur, Jasminka Dobša
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, Varaždin, Croatia
{mbuhin, jasminka.dobsa}@foi.unizg.hr

Luka Kronegger
University of Ljubljana
Faculty of Social Sciences
Kardeljeva ploščad 5, Ljubljana, Slovenia
luka.kronegger@fdv.uni-lj.si

Abstract. *Topic modelling is one of the most popular topics investigated in the area of Natural Language Processing. One of the techniques used for topics modelling is Latent Dirichlet Allocation (LDA). It is an unsupervised machine learning technique which creates topics using a collection of documents based on words or n-grams with similar meaning.*
In this paper, we applied a Structural Topic Model with LDA to extract topics from scientific papers in Social Science. A structural topic modelling of 3663 articles from Web of Science Core Collection from 1999 to 2019 was conducted. The obtained results indicate that an optimal number of topics coincides with the existing number of research areas defined in Social Science or with its integer multiple. This opens an area for research into the comparison between the existing taxonomy and the taxonomy proposed by the LDA model and for the future identification of interdisciplinarity.

Keywords. topic modelling, Latent Dirichlet Allocation, Structural Topic Model, social sciences

1 Introduction

In the last few years, the demand for interdisciplinarity between scientific disciplines has increased. Many interdisciplinary programs are being developed to solve key problems that a single discipline cannot. Interdisciplinarity is not only important in the academic world but also in other areas where it brings about innovation.

In the literature on higher education, interdisciplinary research is defined as "a process of answering a question, solving a problem or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline and draws on the disciplines with the goal of integrating their insights to construct a more comprehensive understanding" (Repko, 2008). In scientometrics, research interdisciplinarity is quantified by examining the network of citations and measuring, for instance, the percentage of citations outside the main discipline of the citing paper.

The automatic identification of interdisciplinarity from a text has already been attempted with text mining approaches (Ramage & Manning & Dumais, 2011), (Chuang et al., 2012), (Nichols, 2014). Dietz et al. (Dietz & Bickel & Scheffer, 2007) used the Latent Dirichlet Allocation (LDA) for topic modelling to quantify the impact that research papers have on each other. Gerrish and Blei (Gerrish & Blei, 2010) showed that LDA could identify a qualitatively different set of relevant articles when compared to traditional citation-count metrics. In the same way, Hall et al. (Hall & Jurafsky & Manning, 2008) identified different methodological trends in the field of computational linguistics across almost 30 years of publications. Despite these studies, which generally utilized LDA as a corpus exploration method, it did not determine the LDA's reliability for recognizing interdisciplinary works. Nichols (Nichols, 2014) presented a novel method for measuring interdisciplinary research in National Science Foundation award portfolios. It proposed using the National Science Foundation (NSF) topic model and the NSF's institutional structure by examining research grant proposals and awards rather than publications. Nanni et al. (Nanni & Dietz & Ponzetto, 2018) investigated the performance of LDA with the outcomes obtained by using other text mining methods such as lexical features within a support vector machine (SVM) or a Rocchio classifier for automatic identification of interdisciplinary works from a corpus of doctoral dissertation abstracts. Considering that, we intend to verify the usefulness of topic modelling for identifying interdisciplinarity in articles. In future work, we would compare the results of the LDA to the results obtained by Social Network Analysis (SNA).

The ultimate goal of the presented analysis is to investigate whether text mining methods, such as Latent Dirichlet Allocation (LDA) topic modelling, could represent a valid alternative for researcher's interest in identifying interdisciplinary fields directly from the textual content of papers titles, abstracts, or keywords. The analysis aims to show how terms characterize a particular scientific field and create new topics using LDA topic modelling. To find the convenient number of topics, we have trained a few topic models using different numbers of topics and

*This paper is published and available in Croatian language at: http://ceciis.foi.hr

evaluated them with measures of semantic coherence of the topics, the likelihood for held-out datasets, residuals, and lower bound on the marginal likelihood.

The paper is organized as follows. The following section describes methods used for topic modelling. Section 3 presents data descriptions and preprocessing techniques, the results of an experiment data analysis, followed by the description of the results. The paper concludes with section 4 related to future work. The research design is presented in Fig. 1.

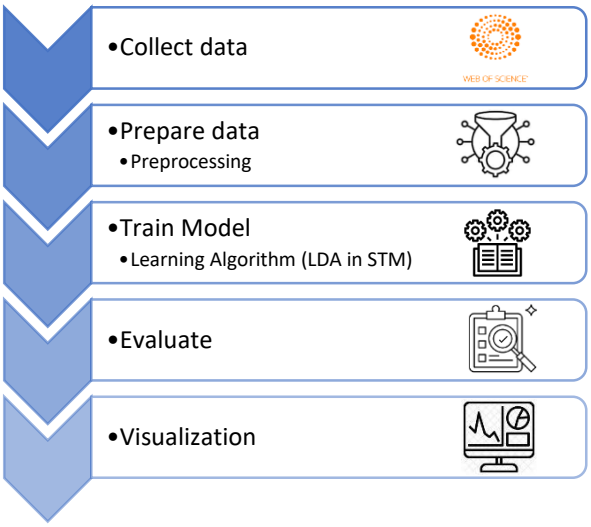

Figure 1. Graphical representation of the research design.

2 Methods

2.1 Text Mining

Text mining aims to detect relevant knowledge that is possibly unknown or covered underneath the obvious one. There are several typical unsupervised and supervised text mining techniques such as text categorization, text clustering, document summarization, and keyword extraction. Topic modelling is a text mining technique which utilizes supervised and unsupervised machine learning techniques.

2.2 Latent Dirichlet Allocation Topic Modelling

Topic modelling is a statistical method which aims to discover an abstract "topics" in a set of documents. It is an unsupervised machine learning technique because it does not require a training dataset or a predefined list of topics. Topics are created from different documents based on words or expressions with similar meaning. One of the most popular methods of topic modelling is Latent Dirichlet Allocation (LDA). LDA attempts to organize all documents to the topics in such a way that the latent topics primarily define the words.

The Latent Dirichlet Allocation method for fitting a topic modelling treats each document as a mixture of topics and each topic as a mixture of words. This method allows documents to "overlap" with each other in terms of content, rather than be separated into discrete groups, which in a way mirrors the typical use of natural language ("Text Mining with R", 2020). LDA is widely used in numerous machine learning, natural language processing (NLP), and information retrieval applications. Griffiths and Steyvers (Griffiths & Steyvers, 2004) used LDA for capturing scientific topics in a collection of documents.

LDA is a generative, probabilistic hierarchical Bayesian model that induces topics from a document collection in three steps (Blei & Ng & Jordan, 2003) (Fig. 2):

1. Each document in the collection is distributed over topics that are sampled for that document based on the Dirichlet distribution.
2. Each word in the document is connected with one single topic based on chosen Dirichlet distribution.
3. Each topic is signified as a multinomial distribution over words that are assigned to the sampled topic.

The following notations will be used:

- M – number of documents,
- N – number of words in each document,
- $\boldsymbol{w}$ – representation of a document represented as a unit-basis vector $\boldsymbol{w} = (w_1, w_2, \dots, w_N)$, where w_n is the n^{th} word in the sequence; vector $\boldsymbol{w}$ has a single component equal to one and all other components equal to zero,
- V – the size of the vocabulary where the v^{th} word in the vocabulary is represented by V – vector such that $w^v = 1$ and $w^u = 0$ for $u \neq v$,
- D – corpus (collection) of M documents, represented as $D = (\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M)$
- k – number of topics a document belongs to,
- z – topic from a set of k topics.

The probability of the observed dataset is calculated and obtained from the corpus D as follows:

$$P(D|\alpha,\beta) = \\ = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d \qquad (1)$$

where variables θ_d are document-level variables, sampled once per document, and variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

There is a parameter α which has the topic distribution θ for each document (Fig. 2). Each of M documents has some θ distribution. θ is a randomly $(M \times k)$ shaped matrix where $\theta(i,j)$ represents the probability of the i^{th} document containing words belonging to the j^{th} topic. θ has a Dirichlet distribution Dir(α).

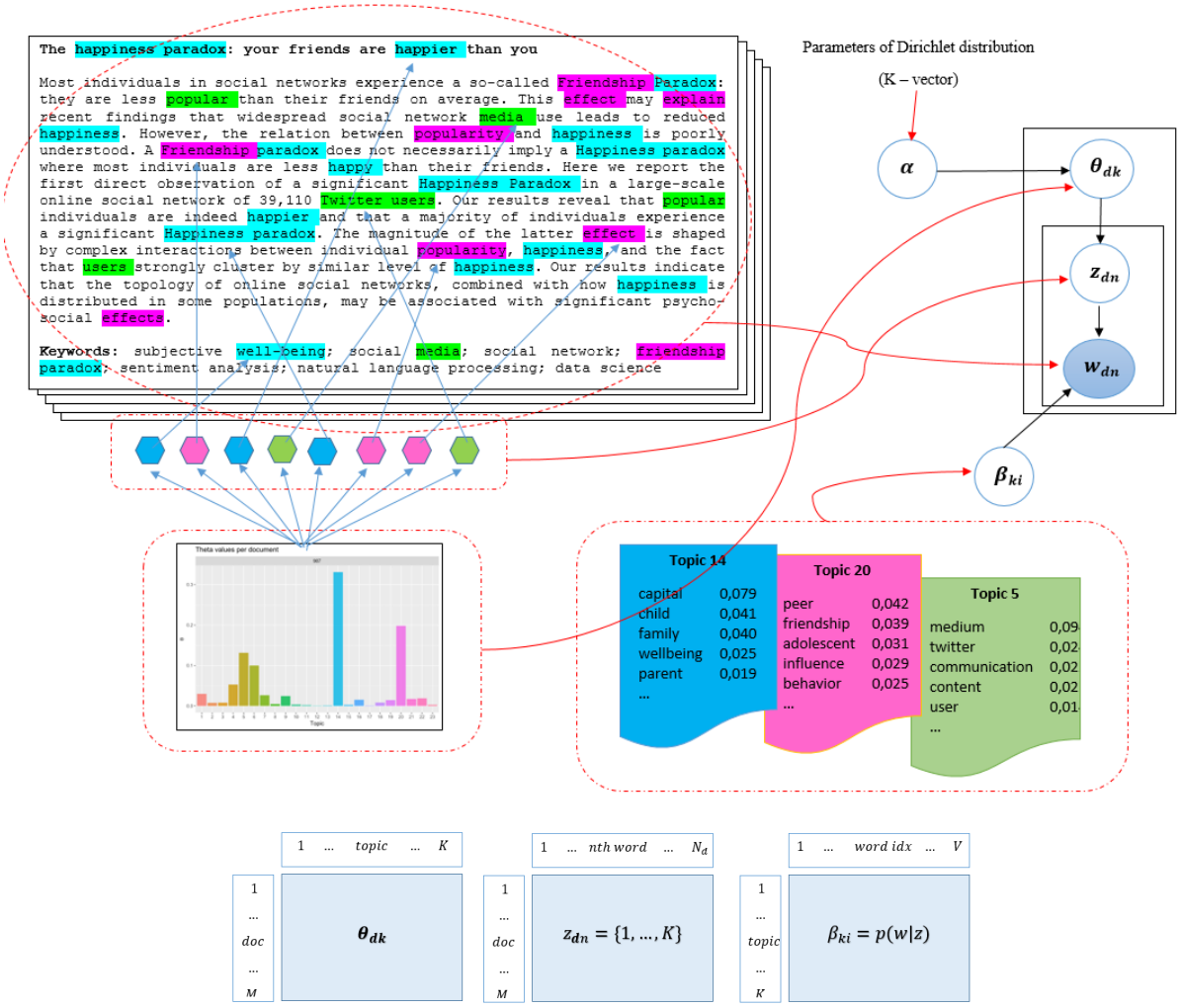

Figure 2. Graphical representation of LDA topic modelling presented on one of the documents from a corpus

Assume that there is a single document with N words, and each word is generated by a topic. We generated N topics that should be filled with words. Based on a single scalar parameter η for each topic, β also has a Dirichlet distribution. β generates k individual words for each topic according to the Dirichlet distribution. Similarly, β is a $(k \times V)$ shaped matrix, where $\beta(i,j)$ represents the probability of the i^{th} topic belonging to the j^{th} word.

To train the LDA model, we need to estimate the hidden parameters α, θ, η and β, where α is a distribution-related parameter that determines what the distribution of topics looks like for all documents in the corpus, θ is a random matrix where $\theta(i,j)$ represents the probability of the i^{th} document containing the j^{th} topic, η is a parameter related to the distribution which determines how words are distributed in each topic, and β is a random matrix where $\beta(i,j)$ represents the probability of i^{th} topic containing the j^{th} word. LDA is a probabilistic model, so we need to calculate the joint distribution of the topic mixture θ, $P(\theta, z, \beta | D; \alpha, \eta)$. For a set of M documents, where each document has N words, and each word is generated by a single topic from a set of k topics, we have to look for the posterior joint probability of θ, z and β, given D and using parameters α and η. The solution to this problem is given in the paper of Blei et al. (Blei & Ng & Jordan, 2003).

For the application of LDA topic modelling, we have used Structural Topic Models (STM) which have an implementation in the **stm** R package ("stm: R Package for Structural Topic Models", 2020).

2.3 Measures for Model Evaluation

The measures we used to find a convenient number of topics are described below.

Semantic coherence is a measure introduced by Mimno et al. (Mimno et al., 2011). The semantic coherence for the topic k is calculated using a list of the N most probable terms in topic k as:

$$C_k = \sum_{n=2}^{N} \sum_{m=1}^{n-1} \log\left(\frac{D(v_n, v_m) + 1}{D(v_m)}\right) \quad (2)$$

where $\boldsymbol{v}$ is a vector of the top N terms in the topic arranged in a descending order, $D(v)$ is the number of documents with at least one term v, and $D(v, v')$ is the number of times that terms v and v' appear together in a document. Intuitively, this is a sum over all term pairs in the top topic terms, returning the log of the co-occurrence frequency divided by the baseline frequency. The one is added in the nominator to prevent taking the log of zero in case a pair of terms never co-occurs. Semantic coherence is maximized when the terms with the highest probability in a given topic frequently co-occur together, and it is a metric that correlates well with the human judgment of topic quality. If there are a few topics that dominate with very prevalent terms, then it is necessary to look at both semantic coherence and exclusivity.

Exclusivity measures the difference between topics by comparing the similarities of word distribution β in various topics. A topic is exclusive if the top words cannot exist among other topics. Exclusivity for term v in topic k is defined as:

$$EX_{k,v} = \frac{\beta_{k,v}}{\sum_{j=1}^{K} \beta_{j,v}} \tag{3}$$

Frequency and exclusivity are important factors in the determination of a term's semantic content. Thus the univariate measure of topical importance may be a useful approximation for different tasks such as dimensionality reduction, feature selection, and content discovery. Therefore, it is modified harmonic mean to move the "average" rank to the lower score (Bischof & Airoldi, 2012). Frequency Exclusivity (FREX) labelling metric is the weighted harmonic mean of the term's rank given as:

$$FREX_{k,v} = \left(\frac{w}{ECDF(EX_{k,v})} + \frac{1-w}{ECDF(\beta_{k,v})} \right)^{-1} \tag{4}$$

where ECDF is an empirical cumulative distribution function for the term v in its topic distribution β_k, and w is a weight for exclusivity set to 0.7 in our experiments.

Another measure for model comparison used to check how well each model predicts terms within the document is the held-out likelihood estimation. In the **stm** package, the held-out likelihood estimation uses two functions ("stm: R Package for Structural Topic Models", 2020). The first function uses the document-completion held-out likelihood method, which is the estimation of the probability of the words occurring within a document when those words have been removed from the document in the estimation step. The second function evaluates the held-out likelihood for missing words based on the model run on the held-out documents. The held-out likelihood estimation is similar to cross-validation and helps to estimate the model's prediction performance.

The assumptions of the model could be tested through residuals. The function `residuals` in the **stm** package measures whether there is overdispersion of the variance of the multinomial variance within the LDA method of generating data. As mentioned in Taddy (Taddy, 2012), if residuals are overdispersed, more topics might be required to absorb some extra variance. Although there is no certain method for choosing the number of topics, both the residuals check and held-out likelihood estimation are useful measures of the number of topics to be selected.

The lower bound is a measure of convergence of the model. Once the bound has a small enough change between iterations, the model is considered converged.

3 Experimental Results

3.1 Dataset and Preprocessing

For experimental evaluation, the analyzed dataset was obtained from the Web of Science (WoS) Core Collection database by searching articles containing phrase *social network** in the WoS Social Science research area in the period from 1999 to 2019. The phrase *social network** is used for the purpose of narrowing of the monitored set of data. The search was performed in March 2020, and a total of 3,664 articles were retrieved. Each of these articles is described with a series of metadata such as author(s), title, abstract, authors' keywords, research area, and year of publication. The main idea is to investigate research topics in the field of social sciences through the number of terms from titles, abstracts, and authors' keywords.

According to the WoS classification, there are 25 categories in the Social Sciences research area. All articles are classified in at least one category from the WoS Social Sciences research area. Only two categories, *Archaeology* and *Development studies*, did not contain any articles from our dataset. Most articles are categorized in *Biomedical Social Sciences*, *Business and Economics*, *Mathematical Methods in Social Sciences*, *Psychology*, *Social Sciences – Other topics* and *Sociology*. Titles, abstracts, and authors' keywords were extracted from each article and merged to get a dataset of one variable and 3,664 instances. 29,199 index terms indexed those articles.

Before the analysis, the dataset was edited using **tm** R package to remove stop words, punctuations, numbers, unnecessary and whitespace characters. High frequently words *social*, *network*, *study*, *analysis*, *model* and *datum* were also removed from the vector of words to reduce the negative impact in the analysis and lemmatization was performed. We created the document-term matrix in which rows represented the documents, and the columns the terms from documents. After described the preprocessing steps, the document-term matrix, created in such a way, had 3,663 documents and 20,718 terms. To reduce the sparsity of matrix, we ejected the index terms

appearing in only one document, and the resulting document-term matrix. After this, the last preprocessing step contained 3,663 documents and 9,096 terms. Finally, each entry in the matrix took values from Term Frequency – Inversed Document Frequency (TF−IDF). Term frequency (TF) is a measure of the importance of a term in a document. Fig. 3 shows the bar chart plot of terms that occur most frequently in the corpus of 3,663 documents. A term's inverse document frequency (IDF) is a measure that penalizes the commonly used terms. By combining IDF with TF measures using multiplication, we obtained the TF-IDF measure of the importance of a term in a document of a corpus. TF-IDF score for the term t in the document d from the document set D was computed as follows:

$$TF - IDF = TF \cdot IDF \tag{5}$$

where

$$TF = \log(1 + freq(t, d)) \tag{6}$$

$$IDF = \log\left(\frac{M}{count(d \in D: t \in d)}\right) \tag{7}$$

where M is the total number of documents and $freq(t, d)$ is the frequency of term t in document d .

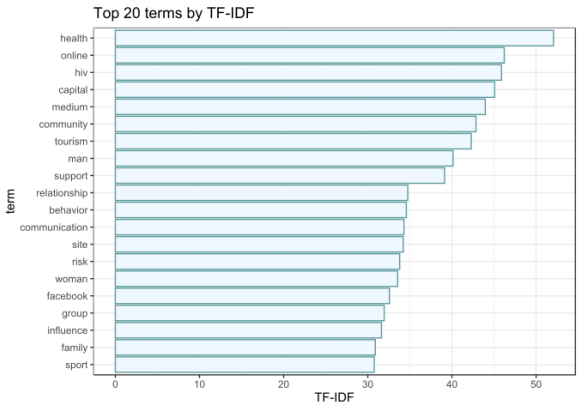

Figure 3. Bar chart plot of the terms with the greatest TF-IDF weight in the dataset.

3.2 Data Analysis

In the next step, we built the Document Term Feature from the corpus and applied the LDA. Before estimating the LDA, we needed to define the number of topics. We trained a group of topic model with a different number of topics and evaluated these topic models to estimate how many topics were appropriate for the given corpus. After we set different values for the number of topics (k) from 2 to 100, we explored how many topics are suitable. These values were intuitively taken to fit the model since there are 25 categories of Social Sciences research area from the WoS classification.

The LDA has two approaches to explore the topics that are estimated. The first approach is to look at how words are associated with topics, and the second approach is to examine documents that are estimated to be highly related to the specific topic.

We evaluated model with measures of semantic coherence of the topics, the likelihood for held-out datasets, residuals and lower bound by making some diagnostic plots to understand how the models perform for the different number of topics and to choose a target number of topics. The topics are then compared with centroids of WoS categories by using cosine similarity. Centroids for specific WoS category k is given by:

$$Cent_k = \frac{1}{n_k}\sum_{j=1}^{n_k} \vec{d}_{k,j} \tag{8}$$

where $\vec{d}_{k,j}$ is the representation of document j in the WoS category k, and n_k is the number of documents in WoS category k. Cosine similarity measures similarity by calculating the cosine of the angle of two vectors $\vec{a}$ and $\vec{b}$:

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \cdot ||\vec{b}||} \tag{9}$$

The similarity between topics obtained by LDA and WoS categories is measured as cosine similarity between the vectors of the word probability distribution of topics and centroids for certain WoS category. The values of the cosine of angle are between 0 and 1 since both vectors have positive values of its elements. We considered that topics from topic modelling are similar to the category from WoS if the value of cosine similarity is greater than 0.5.

3.3 Results

After the model evaluation with semantic coherence of the topics, the held-out likelihood, residuals and lower bound we made some analytical plots using these amounts to know how models perform on a range of topics (Fig. 4). From diagnostic plots, we could see that a good number of topics would be around 25 since around that value growth/fall of corresponding measures of evaluation slows down. When we looked at both semantic coherence and exclusivity of terms to topics together, we could assume that a good choice of the number of topics was 23 (Fig. 5).

The following results are described in two ways. The first approach is to look at the sets of words that are joint with topics. The second approach is to look at the real documents that are estimated to be highly related to each topic. Both of these approaches are presented in Fig. 6. Among the 23 most prevalent topics, topic 22 has the most documents. We can also see that several topics are focused to health (Topic 17 with the most probable terms *HIV*, *man*, *sex*, *sexual*,

risk; Topic 16 with the most probable terms *health, support, age, old, adult*; Topic 15 with the most probable terms *health, care, support, service, access*), communication, internet and social networks (Topic 5 with the most probable terms *medium, twitter, communication, content, user*; Topic 4 with the most probable terms *online, site, internet, Facebook, sns*), tourism, political themes, or business and economics.

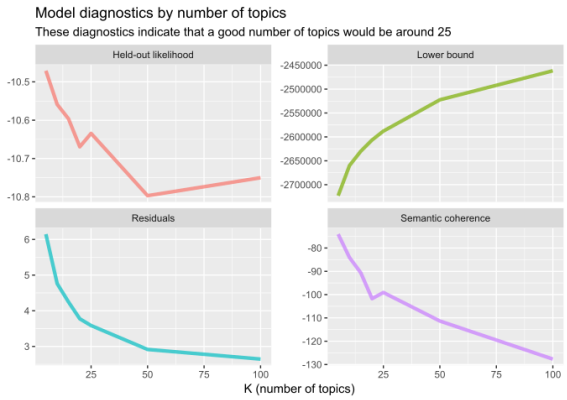

Figure 4. Model diagnostics by the number of topics indicates that a convenient number of topics is around 25.

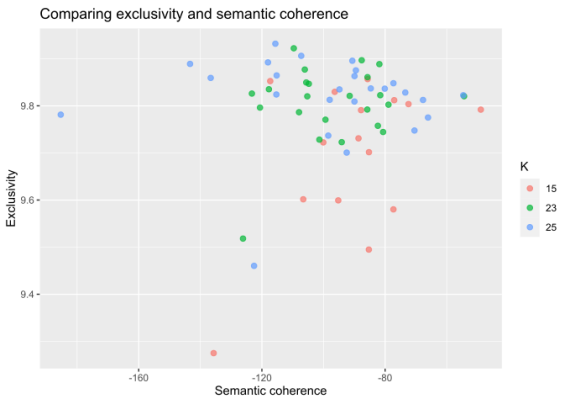

Figure 5. Comparing semantic coherence and exclusivity.

Cosine similarity between vectors based on word probability distribution from topics and centroids for chosen WoS categories (BE – *Business and Economics*, BSS – *Biomedical Social Sciences*, COM – *Communication*, Edu – *Education and Educational Research*, FS – *Family Studies*, GL – *Government and Law*, MathM – *Mathematical Methods in Social Sciences*, Psy – *Psychology*, SI – *Social Issues*, Soc – *Sociology*, SS – *Social Sciences – Other Topics*) is presented in Table 1.

From the results in Table 1, we can see that some topics have something in common with the categories from WoS. For example, Topic 7 is related to the *Mathematical Methods in Social Sciences*, Topic 17 to *Biomedical Social Sciences*, and Topic 11 to *Business and Economics*.

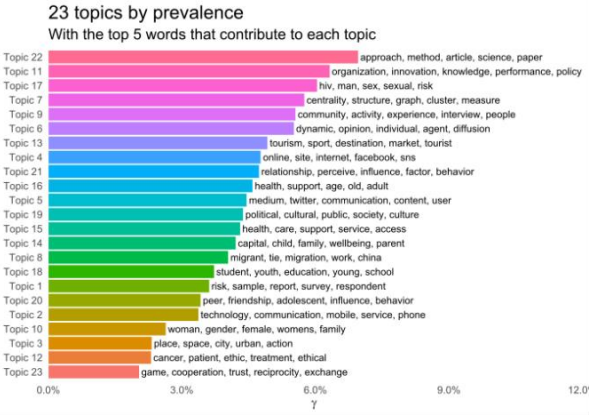

Figure 6. The prevalence of the 23 topics within the entire corpus and top five words associated with the topic.

Table 1. Cosine similarity between topics and selected WoS categories

	BE	BSS	COM	Edu	FS	GL	MathM	Psy	SI	Soc	SS
1	0.17	0.34	0.10	0.07	0.16	0.13	0.22	0.30	0.20	0.24	0.24
2	0.23	0.16	0.24	0.11	0.14	0.13	0.20	0.19	0.29	0.18	0.31
3	0.20	0.20	0.09	0.21	0.16	0.21	0.20	0.22	0.22	0.28	0.29
4	0.21	0.16	0.32	0.11	0.13	0.20	0.14	0.24	0.29	0.19	0.36
5	0.21	0.13	0.26	0.13	0.11	0.18	0.18	0.20	0.17	0.15	0.32
6	0.36	0.25	0.10	0.16	0.21	0.22	0.54	0.27	0.20	0.35	0.37
7	0.40	0.23	0.08	0.17	0.16	0.13	0.68	0.28	0.17	0.43	0.33
8	0.32	0.26	0.10	0.16	0.31	0.30	0.26	0.30	0.23	0.35	0.35
9	0.20	0.28	0.11	0.22	0.21	0.18	0.18	0.30	0.22	0.29	0.32
10	0.13	0.29	0.06	0.13	0.33	0.15	0.11	0.24	0.15	0.18	0.19
11	0.56	0.24	0.12	0.23	0.16	0.27	0.37	0.29	0.26	0.35	0.45
12	0.21	0.35	0.08	0.14	0.12	0.18	0.17	0.32	0.29	0.22	0.26
13	0.46	0.15	0.12	0.20	0.14	0.16	0.18	0.19	0.18	0.30	0.42
14	0.22	0.27	0.08	0.22	0.19	0.15	0.15	0.30	0.26	0.37	0.31
15	0.15	0.49	0.06	0.13	0.16	0.14	0.13	0.33	0.24	0.29	0.24
16	0.17	0.50	0.07	0.17	0.16	0.14	0.17	0.38	0.22	0.35	0.27
17	0.14	0.60	0.09	0.07	0.47	0.10	0.14	0.50	0.17	0.19	0.22
18	0.20	0.20	0.14	0.44	0.16	0.19	0.17	0.32	0.30	0.32	0.33
19	0.23	0.18	0.17	0.22	0.17	0.37	0.20	0.26	0.25	0.26	0.36
20	0.27	0.29	0.19	0.12	0.15	0.16	0.35	0.30	0.23	0.33	0.28
21	0.26	0.27	0.12	0.13	0.16	0.16	0.22	0.32	0.27	0.29	0.30
22	0.37	0.23	0.12	0.29	0.20	0.24	0.37	0.27	0.24	0.34	0.42
23	0.31	0.24	0.10	0.27	0.21	0.13	0.32	0.30	0.29	0.30	0.30

We can also notice that some topics do not overlap with only one category from WoS but several, so we can assume that there is interdisciplinarity between these scientific disciplines from WoS, which is not surprising, given that most articles are categorized into multiple disciplines. Topic 22 has approximately equal values of cosine similarity for *Business and Economics*, *Mathematical Methods in Social Sciences*, *Sociology* and *Social Sciences – Other Topics*, while topic 17 has approximately equal values of cosine similarity for *Biomedical Social Sciences*, *Family Studies*, and *Psychology*. Another thing we noticed is

that individual areas from the WoS categories have approximately equal cosine similarity values for different topics. We can assume that this is because the LDA algorithm identified subcategories within this category. For example, *Sociology* has approximately equal cosine similarity values for topics 6, 7, 8, 11, 14, 16, 20 and 22.

4 Conclusion

In this paper, we have given a brief introduction of Latent Dirichlet Allocation (LDA) topic modelling, which was applied in Structural Topic Models (STM) on the dataset from the Web of Science Core Collection.

The main goal of the research was to compare topics obtained by LDA topic modelling with categories of Web of Science Core Collection for the field of Social Sciences. The research was conducted on the sample of papers from 1999 to 2019 with the keyword *social network** and the results are restricted to publications containing that phrase. The comparison between the topic obtained by LDA and the given taxonomy in terms of cosine similarity indicates that social networks are mainly applied in disciplines of *Business and Economics* (BE), *Biomedical Social Sciences* (BSS), *Mathematical Methods in Social Sciences* (MathM) and *Psychology* (Psy) which seems like an intuitive result. Furthermore, based on cosine similarities, we were also able to identify interdisciplinarity between disciplines of BE and MathM, BSS and MathM, BSS and Psy, BSS, FS, and Psy.

Based on the intuitive results obtained on this sample of papers, we plan to extend our research to all papers in the collection in the field of Social Sciences to identify interdisciplinary fields and conduct further research by analyzing social networks and symbolic data.

In future research, we intend to investigate the interdisciplinarity between science disciplines which are hidden or masked and reconsider the existing taxonomy of research areas in social sciences and its temporal changes.

References

Bischof, J., Airoldi, E. (2012). Summarizing Topical Content with Word Frequency and Exclusivity. *Proceedings of the 29th International Conference on Machine Learning, ICML '12* (pp. 201–208). New York: J. Langford, J. Pineau (eds.).

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, pp. 993-1022.

Chuang, J., Ramage, D., Manning, C., Heer, J. (2012). Interpretation and trust: designing model-driven visualizations for text analysis. *SIGCHI Conference on Human Factors in Computing Systems*, (pp. 443-452). Austin, Texas, USA.

Dietz, L., Bickel, S., Scheffer, T. (2007). Unsupervised prediction of citation influences. *24th international conference on Machine learning* (pp. 233-240). Corvalis, Oregon, USA: Association for Computing Machinery, New York, United States.

Gerrish, S., Blei, D. (2010). A Language-based Approach to Measuring Scholarly Impact. *27th International Conference on Machine Learning*, (pp. 375-382). Haifa, Israel.

Griffiths, T., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, pp. 5228-5235.

Hall, D., Jurafsky, D., Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. *Conference on Empirical Methods in Natural Language Processing*, (pp. 363–371).

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A. (2011). Optimizing semantic coherence in topic models. *Conference on Empirical Methods in Natural Language Processing (EMNLP '11)* (pp. 262-272). USA: Association for Computational Linguistics.

Nanni, F., Dietz, L., Ponzetto, S. P. (2018). Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. *Digital Scholarship in the Humanities, Volume 33, Issue 3*, pp. 612–620.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics 100(3)*, pp. 741-754.

Ramage D., Manning C. D., Dumais S. (2011). Partially labelled topic models for interpretable text mining. *17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457-465). San Diego California USA: Association for Computing Machinery New York NY United States.

Repko, A. F. (2008). *Interdisciplinary Research: Process and Theory.* California: Sage: Thousand Oaks.

Roberts, M. E., Stewart, B. M., Tingley, D. (August 2020). *stm: R Package for Structural Topic Models.* Retrieved from The Comprehensive R Archive Network: http://www.structuraltopicmodel.com/

Silge, J., Robinson, D. (August 2020). *Text Mining with R.* Retrieved from https://www.tidytextmining.com/topicmodeling.html

Taddy, M. A. (2012). On Estimation and Selection for Topic Models. *The 15th International Conference on Artificial Intelligence and Statistics.*, (pp. 1184-1193).