Universiteit Utrecht

Developing text-mining methods to review the published literature

Rania Koutsoukou Prelorentzou

Applied Data Science MSc thesis

Utrecht University

Under the supervision of Dr. Caspar J. van Lissa

Faculty of Social and Behavioral Sciences, Utrecht University

Second supervisor: Rens van de Schoot

July 2022

Author Note

Student ID: 0711454, Email: o.koutsoukouprelorentzou@students.uu.nl

Abstract

Text mining is considered an effective approach for the identification of relevant phenomena in systematic reviews. Topic models have shown to be a promising unsupervised technique to reveal common topics in text data. This research used three topic modeling text mining algorithms, *LDA*, *Top2Vec*, and *BERTopic*, to identify the relevant phenomena in two datasets from published literature text data. The first dataset contains bibliographic data of articles about adolescents' emotional regulation, and the second, bibliographic data of articles about cooperation in prisoner's dilemma, where each of the datasets is divided to abstracts and keywords. The goal of this thesis is to select the optimal number of topics/phenomena and then map them to a network. Comparing the performance of the three algorithms with regards to topic quality and network representation of the topics, it is concluded that *BERTopic* produced more meaningful topics than *Top2Vec* and *LDA*. The code is provided at: https://github.com/raniakp/thesis-text-mining-published-literature

Keywords:

text mining systematic review, phenomena identification, LDA, Top2Vec, BERTopic, topic quality

Contents

Introduction

Theory construction is the process of developing and combining heterogeneous theories into coherent unities, or the process of modifying and expanding theories in the light of logical, semantic, and empirical analysis (Markovsky and Webster Jr, 2007). The methodology of theory construction requires a sequence of steps, including identification of empirical phenomena, development of a prototheory, model construction to examine the prototheory, and evaluation of the process (Herfel, 1995, Friedman, 2003, Borsboom et al., 2021). Phenomena are stable features of the world that scientists aim to explain (Bogen and Woodward, 1988, Haig, 2014). Prior to their explanation, identification of these phenomena must be conducted. So far, the identification process is based on the experience and empirical knowledge of the scientist (Haig, 2013). However, in the past few years, research has been done in order to automate this process. Text mining systematic reviews have been suggested as an objective tool for detecting phenomena (van Lissa, 2021, Usai et al., 2018, Li et al., 2016, O'Mara-Eves et al., 2015, Thomas et al., 2011).

Text mining is a field that has increased rapidly in recent years. With the enormous amount of text data available from various applications, innovations in algorithmic design are required to learn meaningful patterns from the data in a dynamic and scalable manner (Aggarwal and Zhai, 2012). Thus, mining techniques concentrate on the most important models, algorithms, and applications for determining what can be learned from various types of text data. In addition, unstructured data is used in many text mining algorithms, such as text clustering, text categorization, topic tracking, summarization, and recommender systems (Breck et al., 2007, Choi et al., 2005).

Systematic reviews (SR) include the identification, evaluation and synthesis of all relevant studies for specific topics (Li et al., 2016). High-quality SRs adhere to tight guidelines and take a lot of time and work. In order to minimize the effort of abstract filtering in systematic reviews and automate the process of summarizing high-level information, text-mining methods are used.

Given the increased amount of research in this field, lately, studies have been focused on text mining within systematic reviews (Usai et al., 2018, O'Mara-Eves et al., 2015,

Thomas et al., 2011). With regards to research, text mining techniques, such as term extraction, document clustering, document classification and query expansion are used by reviewers aiming to automate these processes that so far are performed manually (O'Mara-Eves et al., 2015, Ananiadou et al., 2009). Three self-defined semantics-based ranking measures, including keyword relevance, indexed-term relevance, and topic relevance, were proposed as part of a text mining SR supporting framework (Li et al., 2016). In addition, research has suggested that text mining systematic review is an effective and rather objective tool for detecting phenomena (van Lissa, 2021). According to the aforementioned study, using published literature for the phenomena detection has been considered an efficient approach. This method suggested that the frequency and the co-occurrence with which phenomena appear in the literature are signs of their relevance. Moreover, Amy van der Ham, research assistant of Dr. Caspar J. van Lissa, did a further analysis in this problem using ASReview (van de Schoot et al., 2021), word2vec (Church, 2017) and GloVe (Pennington et al., 2014) for feature extraction and k-means (Forgy, 1965) and DBSCAN (Ester et al., 1996) algorithms for the clustering.

The current study seeks to explore topic modeling algorithms for relevant phenomena identification. Specifically, the goals of this thesis are to extract features from a dataset of published literature data and select the optimal number of topics/phenomena. Then, the generated topics will be represented as a network, because mapping the relationships between phenomena can show how relevant they are. Thus, the research question can be formulated as follows: Which text mining method identifies the optimal number of topics/phenomena for published literature data?

Regarding the exploration research that has already been done, in this thesis three text mining methods are compared for phenomena detection. First, *LDA* (Blei et al., 2003) is applied as a baseline method to extract the corresponding topics/phenomena. Secondly, I will try to improve upon the baseline by using *Top2Vec* (Angelov, 2020) and *BERTopic* algorithms (Grootendorst, 2022). From published literature text data, two datasets are used, articles about adolescents' emotional regulation, and articles about cooperation in prisoner's dilemma, and from each dataset, abstracts and keywords are analyzed.

Data

In this research, two bibliographic datasets were used, where the first one is a database of 6305 articles on teenagers' emotional problems(van Lissa, 2021), from the Web of Science, and the second one is a database of 2004 articles on cooperation in prisoner's dilemma, from the Cooperation Databank. The teenagers' emotional problems dataset contains 74 variables and 6305 rows, and the cooperation in prisoner's dilemma dataset contains 47 variables and 2004 rows. The most important common variables between the two datasets are title, keywords, abstract, authors' names, year, volume, pages, URL, and type.

From each dataset, two dataframes were created, one with the keywords and one with the abstracts. Firstly, in each dataframe the NA values and the duplicates were excluded, and then keywords and abstracts were extracted by the document. Using Exploratory Data Analysis to make the data suitable for the current research, pre-processing steps were taken, corpus and dictionaries were created, stop words and punctuation were removed, and lemmatization was applied. The resulting corpus were: (i) 'teenagers' emotional problems' keywords of 5024 documents (1281 excluded) with 1492 unique words and 18.25 average number of words in each document, (ii) abstracts of 6087 (218 excluded) with 7888 unique terms and 120.96 average number of words in each document, (iii) the cooperation in prisoner's dilemma's keywords of 1729 documents (315 excluded) with 987 unique terms and 16.78 average number of words in each document, and (iv) abstracts of 2004 documents with 3791 unique terms and 98.04 average number of words in each document.

For further analysis, by using the *'FreqDist'* library (Bird et al., 2008), which allows to determine the count of the most common terms in a corpus, word clouds and barplots were created based on the frequencies of the most common words in each corpus. The word clouds, A1a and A2a, that were created respectively for abstracts and keywords of the "teenagers' emotional problems" dataset, have similar most common words, such as "child", "regulation", "study", "emotional". Moreover, regarding the barplots with the 25 most common word frequencies, A1b and A2b , based on the fact that terms such as

"child", "regulation", "adolescent" and "study" have such a high frequency in the corpus, it might be a good idea to remove them (ie. add them to the stopwords) prior to analysis. The same observation holds true for the words "game", "cooperation", "social", "dilemma", "good" and "group" from cooperation in prisoner's dilemma dataset A3a, A3b, A4a and A4b. These words most likely are not helpful for phenomena identification due to their existence in a large amount of documents.

Methods

Topic modeling is the process of identifying topics in a set of documents. In this section, the three topic modeling algorithms that will be used for phenomena detection will be explained, *LDA*, *Top2Vec*, and *BERTopic*, along with the methods that will be applied to evaluate the models.

LDA

LDA, which stands for *Latent Dirichlet Allocation* (Blei et al., 2003), is a widely used topic modeling algorithm that was introduced as the first approach that enables the modeling of topic semantics entirely within the framework of Bayesian statistics (Maier et al., 2018). It should be noted that this method processes documents as bag-of-words(BOW), which takes into consideration only the frequency of words in a document and not their order. Given that text data contain documents consisting of words, *LDA* assumes that each topic is generated by a combination of words and each document is generated by a combination of topics. As a result, each topic is modeled as a probability distribution over our vocabulary, while each document as a distribution over the topics. Firstly, the number of topics, k , is a hyperparameter that should be defined by the user. Then, *LDA* assigns randomly every word of each document to one of the k topics. By observing only words in the documents, the model estimates the probability of each word belonging to a topic and iteratively updates the aforementioned probability distributions. When the model is trained the converged probability distributions can be used to find which words represent mostly each topic. In this research, the unsupervised method of *LDA* is used for topic modeling, where the extracted topics can be treated as the relevant phenomena.

LDA will be used as a baseline method to extract the corresponding topics and assign the words of the corpus to them. One major limitation of this method is that the number of topics is a hyperparameter that should be manually selected (Maier et al., 2018). Thus, to find the optimal number of topics, the model will be trained for various numbers of topics. Then, the optimal number will be chosen based on the coherence and diversity of the produced topics (Dieng et al., 2020, Röder et al., 2015). Further explanation on the evaluation of these metrics is given in subsection Evaluation.

Given that in the aforementioned method, firstly the order of words is not taken into consideration, secondly the number of topics should be defined, and thirdly it requires some preprocessing steps, such as lemmatization, stemming and custom stop-words lists (Blei et al., 2003), we can not be sure that the representations of the words will be sufficient to extract the clusters. Consequently, two other state-of-the-art algorithms will be used too in this research, namely *Top2Vec* and *BERTopic*.

Top2Vec

Currently, in the field of Natural Language Processing (NLP) distributed representations for words and documents have gained more popularity due to their ability to capture their semantics. A distributed representation of any concept is defined as a representation that is distributed over various processing units. The introduction of the new algorithm, *Top2Vec* is based on two ideas, firstly, the creation of a space with words and documents representations, and secondly, the distributional hypothesis, which implies that words with similar meanings are used in similar contexts (Angelov, 2020). Specifically, for *Top2Vec*, in a combined embedded document and word space, a dense area can be understood as a similar topic.

Regarding *Top2Vec* algorithm, the first step is to create a joint document and word embedding space. In order for this to be achieved, the distance between document vectors and word vectors is interpreted as a semantic association, which means a direct or indirect relation between two entities (words in this case) that is considered meaningful. Hence, documents with semantic similarity should be placed closely, and words should be near documents that they best describe. This spatial representation is called a semantic space (Griffiths et al., 2007), where topic vectors can be calculated. To create this space, *doc2vec* Distributed Bag of Words (DBOW) (Le and Mikolov, 2014) architecture is used (Lau and Baldwin, 2016), which is an extension of *word2vec* (Mikolov et al., 2013). So, in the semantic space, the word vectors, which are the most semantically relevant to the document's topic, are those that are closest to the document vector. In addition, a dense area of documents in this space is an area with highly similar documents. Thus, this dense area can be understood as a topic with common documents (Angelov, 2020).

Having created the semantic space, the next step should have been to identify directly the topics. However, the high-dimensionality of the generated document and word vectors, regularly of 300 dimensions, arises two issues. Firstly, the document vectors are very sparse in space, and secondly, the computational cost is high due to this sparsity. The problem with processing high-dimensional vectors is called "curse of dimensionality" (Indyk and Motwani, 1998). To further explain, the distance gets larger as dimensionality increases, since each new dimension adds a new non-negative term to the sum used to calculate the euclidean distance. Moreover, the distance to the nearest point tends to approximate the distance to the farthest point, which leads to unclear spatial locality as it is harder to distinguish neighborhoods of points that lie close to each other and far from other neighborhoods Aggarwal et al., 2001. In order to overcome the "curse of dimensionality", in *Top2Vec*, dimension reduction is computed with the algorithm Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes et al., 2018). UMAP was chosen for *Top2Vec* compared to other dimension reduction algorithms, such T-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008), because it preserves local and global structure, and is able to scale to very large datasets.

After the dimensionality reduction, the method of Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes and Healy, 2017, Campello et al., 2013) is used on document vectors to identify the dense areas. HDBSCAN method combined with UMAP-reduced dimension not only recognises the dense areas of documents but also the noise documents. Basically, it labels each document of the semantic embedding space with the name of the dense area that it belongs or as noise. Then, to find the topic vectors the method of calculating the centroid is chosen. As a result, every point in the semantic space represents a topic, whose semantics are best defined by the nearest word vectors. The semantic similarity of each word vector to the topic vector is determined by the distance between them. Consequently, stop-words are rarely seen close to the topic vectors because they are recognised as noise, therefore stop-word removal as a pre-processing step is unnecessary in this algorithm(Angelov, 2020).

All in all, *Top2Vec* is an unsupervised learning algorithm that finds topic vectors in a semantic space, which consists of jointly embedded document and word vectors. So, unlike traditional topic modeling techniques, using this algorithm can lead us to directly define the number of topics/phenomena of our datasets and specify the most relevant words of each topic.

BERTopic

The third algorithm that will be used in this research is *BERTopic*, which is similar to *Top2Vec* in terms of algorithm structure. Firstly, *BERTopic* creates document embeddings using pre-trained transformer-based language models. To be more specific, the rapidly growing state-of-the-art model, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), is used by *BERTopic* as a text embedding technique. Then, the algorithm clusters these embeddings, and in the end, generates topic representations with the *class-based TF-IDF* process (Grootendorst, 2022). Further information on this process will be given in this subsection.

To begin with the explanation of the aforementioned process, *BERTopic* assumes that semantically similar documents contain the same topic. Therefore, *Sentence-BERT (SBERT)* (Reimers and Gurevych, 2019) framework is used to create document embeddings in vector space. Based on their context, *SBERT* converts sentences and paragraphs to dense vector representations using pre-trained language models. For instance, the same sentence can have a different embedding representation based on the context of the surrounding sentences. This process can also be achieved by using other embedding techniques if fine-tuning in semantic similarity is performed. Consequently, as new language models are produced, the quality of clustering in *BERTopic* will be improved (Grootendorst, 2022).

Continuing with *BERTopic*, document clustering should be performed. However, as explained in *Top2Vec*, one limitation of the algorithm is the "curse of dimensionality", because the embedded document vectors are very sparse in space. To overcome this issue, *UMAP* is used again due to its ability to maintain more of the local and global features of high-dimensional data in lower projected dimensions. Additionally, UMAP is applicable

to language models with various dimensional spaces. After the embedding reduction, *HDBSCAN* is used for clustering. As in *Top2Vec*, *HDBSCAN* uses this technique to represent clusters, allowing noise to be modeled as outliers. This eliminates the assignment of unrelated documents to any cluster and is likely to improve topic representations (Grootendorst, 2022).

In the end, the topic representations are based on the documents in each cluster, with one topic given to each cluster. A modification of the classic *TF-IDF* method (Joachims, 1996) is used in order for this to be achieved. In the classic *TF-IDF* model word frequency and inverse document frequency are combined and calculated. In *BERTopic* all documents in a cluster are concatenated and they are treated as one. So, instead of document frequency, inverse cluster/class frequency is calculated and combined with the word frequency. Given that the clustering process has already been achieved by HDBSCAN, inverse cluster/class frequency will be measured in order to specify how much information a word provides to a cluster. Thus, this modified class-based *cTF-IDF* procedure focuses on the words importance instead of documents importance. This leads to generating topic-word distributions for each cluster/class of documents. Moreover, the user can define a value in order for the least common topics to be merged with the most similar ones (Grootendorst, 2022). Last but not least, the final results of *BERTopic* contain n number of topics, and also an extra topic, named -1, with all the most common words of the datasets.

Evaluation

Concerning this thesis, to validate our results *OCTIS (Optimizing and Comparing Topic models is Simple)* (Terragni et al., 2021), an open-source python package, will be used. As already mentioned, three algorithms will be compared, *LDA*, *Top2Vec*, and *BERTopic* for the given datasets. Hence, topic quality will be measured by multiplying two frequently used metrics, topic coherence, and topic diversity (Dieng et al., 2020). These measures are utilized to assess the effectiveness of the topic models in this study. Topic coherence is a quantitative measure that is defined as the average of pairwise word similarities formed by top words of a given topic (Rosner et al., 2014). To be more

specific, in a coherent topic the most likely words should have a high degree of common information. For each topic model, its topic coherence will be evaluated using C_v (Syed and Spruit, 2017, Röder et al., 2015), which according to this research performs better than other topic coherence measures, such as C_{NPMI} or C_{UMass}. The measure ranges from [0,1] where 1 indicates the perfect topic coherence. In addition, topic diversity is the percentage of unique words for all topics, and it ranges from [0, 1] where 0 indicates redundant topics and 1 indicates a wider range of topics (Dieng et al., 2020). So, in this thesis, the product of topic coherence and topic diversity which is defined as the topic quality, will be the evaluating metric of our models.

As it is already mentioned, topics will represent the phenomena in our study. Hence, in order to map the phenomena, a network will be created for each dataset and model. The Python library *Pyvis* [1] and the Python package *NetworkX* [2] will be used to create the networks. Mapping the phenomena in a network, firstly, will add information on the evaluation process, and secondly, will give insights into the connection between the phenomena. For each network, the nodes will be represented as the topics/phenomena, named by the first two words of each topic, and edges will be represented as the existence of common words between the topics. For instance, if two topics contain a common word among their 10 most representative words, an edge will be drawn between those two topic nodes.

Hyperparameter Tuning

For each of the aforementioned models, hyperparameter tuning is needed. In this subsection, the selection of the hyperparameters is explained.

It has already been stated that *LDA* requires the number of topics in advance. Since the identification of phenomena/topics is the main goal of this research, the number of topics could not be predefined. So I run the model for every dataset for 3 to 150 topics and then the number with the highest topic quality was chosen. However, for the emotional regulation dataset the number of topics with the highest topic quality was 3

[1] https://pyvis.readthedocs.io/en/latest/tutorial.html

[2] https://networkx.org/

for both abstracts and keywords, and it can be seen in Figures B1a, and B1b, it is incomprehensible. As a result, I run again the model for every dataset, iterating over the numbers of topics from 7 to 150, where 7 and 150 were considered more logical choices for these datasets, given the results of topic coherence and topic diversity plots from 7 to 150 number of topics, see Figures C1a, C1a, C1c, and C1d. Moreover, *LDA* needs to be provided with the corpus of the given documents, the dictionary with the words of the documents, and the maximum number of iterations which allows *LDA* to converge. Creating the corpus and the dictionary, the words that occur at least in 5 documents were kept, and also a word had to occur in less than 85 percent of the document in order to be included in the dictionary. The number of iterations is set to 1000, since *LDA* converged there. Last but not least, the random state parameter is defined to prevent the randomness of the results.

Regarding *Top2Vec*, for embedding model parameter, I tried *'universal-sentence-encoder'*, *'doc2vec'*, and *'all-MiniLM-L6-v2'*. *'Universal-sentence-encoder'* Cer et al., 2018 was chosen, since it performed slightly better than the other models. In addition, the topic quality was higher when *n_gram* was set to TRUE and the model used bigrams.

For *BERTopic*, *"paraphrase-MiniLM-L3-v2"* [3] was used as the embedding language model. *Paraphrase-MiniLM-L3-v2* is considered the sentence transformers model with the best trade-off of performance and speed when limited GPU capacity is available (Grootendorst, 2022). *BERTopic* has also a parameter which is called *min_topic_size*. The higher this value is, the lower the number of clusters/topics is. The default value of the parameter is 10, but I set it to 15 for the emotional regulation dataset and to 7 for prisoner's dilemma dataset, because the reproduced topics were more logical.

[3]https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L3-v2

Results

In this section, the results of the analysis will be presented. Firstly, the results of keywords analysis from the emotional regulation dataset will be provided, then the abstracts from the same dataset and then the keywords, and abstracts analysis for the prisoner's dilemma dataset.

For the keywords of the emotional regulation dataset, the highest topic quality value is 0.247 for the *LDA* model compared to 0.07 of *Top2Vec*, and 0.216 of *BERTopic*, and also, the number of topics for these values is 20, 40 and 94 topics respectively, as shown in Table 1. By observing the results of topic coherence and topic diversity in Table 1, it seems that the highest topic coherence value belongs to *Top2Vec*, 0.685, compared to 0.422 of *LDA* and 0.371 of *BERTopic*, which implies that the topics in *Top2Vec* are more coherent. However, the topic diversity of *Top2Vec* is 0.105 in contrast to the higher values of the other models. This explained why the topic quality value of *Top2Vec* is significantly lower than the others, Table 1.

Table 1
Topic Coherence and Topic Diversity for emotional regulation dataset

Dataset	Model	# topics	Coherence	Diversity	Quality
Keywords	*LDA*	20	0.422	0.585	**0.247**
	Top2Vec	40	0.685	0.105	0.07
	BERTopic	94	0.371	0.583	0.216
Abstracts	*LDA*	10	0.303	0.52	0.157
	Top2Vec	40	0.864	0.079	0.154
	BERTopic	52	0.419	0.456	**0.191**

In Table 1, the results of abstracts analysis from the emotional regulation dataset are also shown. There, the highest topic quality value is 0.191, recorded for *BERTopic* with 52 topics, 0.419 topic coherence, and 0.456 topic diversity. As in keywords analysis, it is observed that *Top2Vec* has the highest topic coherence value, 0.864, but the topic diversity is 0.079, which again is significantly lower than the topic diversity values of the other models. The low topic diversity means that the different topics do not have many unique words. As for the number of topics, *LDA* and *Top2Vec* have fewer topics than

BERTopic, 20 and 40 topics, respectively.

Considering the prisoner's dilemma dataset, the analysis of the keywords displays that *Top2Vec* has the highest topic quality with 0.230 and 10 topics, compared to 0.07 for *LDA* and 8 topics, and 0.212 for *BERTopic* and 38 topics, as shown in Table 2. It is worth mentioning that *LDA* and *Top2Vec* indicate almost the same number of topics, although they differ a lot in terms of topic quality because *Top2Vec* has approximately 3 times higher value than *LDA*.

Table 2
Topic Coherence and Topic Diversity for prisoner's dilemma dataset

Dataset	Model	# topics	Coherence	Diversity	Quality
Keywords	*LDA*	8	0.218	0.325	0.07
	Top2Vec	10	0.397	0.580	**0.230**
	BERTopic	38	0.416	0.512	0.212
Abstracts	*LDA*	7	0.277	0.642	0.178
	Top2Vec	17	0.846	0.158	0.133
	BERTopic	48	0.360	0.744	**0.267**

The abstracts analysis of prisoner's dilemma dataset demonstrates that *BERTopic* has the highest topic quality value of 0.267, with topic coherence 0.360, and topic diversity 0.158. Specifically, as can be seen in Table 2, this value is almost double than *LDA* and *Top2Vec*. Moreover, noticing the number of topics, *BERTopic* has 48, *LDA* 7, and *Top2Vec* 17. Once again, *Top2Vec* has the highest coherence, 0.846, and the lowest topic diversity, 0.158.

As already mentioned, regarding *Top2Vec*, the values of topic diversity are the lowest in 3 out of 4 cases/datasets. Additionally, the topic coherence values are the highest for the same cases. Further information about these observations will be given through Graphs C2a C2b, D1a, and D1b. In these graphs, the word embeddings are represented as points in a 2-Dimensional space with different colors for each topic. The name of each topic is the first 2 words or bigrams with the highest scores. As shown in graph D1a, the clusters/topics are easily distinguishable compared to the other 3 graphs. This aligns with the aforementioned results from Table 2 for the keywords, where the topic diversity

is approximately 5 times higher than the topic diversity values achieved by *Top2Vec* in the other cases.

For *BERTopic*, the number of topics is considerably higher in every case than in the other models. It is also observed that the values of *BERTopic* have a small range. For topic quality, the values are between 0.267 to 0.191, while in *Top2Vec* they are between 0.230 to 0.07, and in *LDA* between 0.247 to 0.07. These results indicate that *BERTopic* is more stable than the other two.

To better evaluate the results, network graphs are created for each model for both datasets for abstracts and keywords. For readability purposes, the networks are placed at the end of the thesis as an appendix because decreasing their size would result in an unclear demonstration of topics. In Table 3, the links to the figures of the networks are presented. The name of each topic in the networks is the first 2 words or bigrams with the highest scores.

Table 3
For each dataset, the produced networks of each model can be seen in the corresponding links

Dataset	Cases	LDA	Top2vec	BERTopic
Emotion Regulation	*Keywords*	E1	E5	E9
	Abstracts	E2	E6	E10
Prisoner's Dilemma	*Keywords*	E3	E7	E11
	Abstracts	E4	E8	E12

With regards to the networks, the most informative networks seem to be reproduced by *BERTopic*. The topic quality of the keywords for both datasets has shown that *BERTopic* did not perform as well as the other models, despite the fact that the topics of the *BERTopic* networks, E11, E12, E9, E10, represent meaningful topics, and correspond to the topic coherence values of *BERTopic*. Furthermore, these networks are not as dense as *LDA* and *Top2Vec* networks, because the number of edges between the nodes is fewer. This agrees with the values of topic diversity whose range for *BERTopic* is between 0.456 to 0.744, as depicted in Tables 1, and 2.

At last, one example of the produced topics for the abstracts of emotional regulation dataset will be presented. In Tables 4, 5, and 6 the topics of all the models can been seen. The words with the highest scores of each topic are the same with the names of the topics as represented in the networks.

Table 4
The 10 Topics that are produced from LDA with the 4 words with the highest scores for each topic

Topic	Words with highest scores
0	student, adolescent, study, self
1	self, adolescent, disorder, study
2	problems, emotionnal, patient, study
3	adolescent, immigrant, study, regulation
4	teacher, rating, infant, problems
5	problems, adolescent, disorder, study
6	adolescent, problems, self, study
7	family, child, violence, representation
8	regulation, emotion, study, disorder
9	stress, regulation, emotion, study

Table 5
The 40 Topics that are produced from Top2Vec with the 2 words with the highest scores for each topic

Topic	Words with highest scores	Topic	Words with highest scores
0	emotion regulation, emotion	**20**	developmental psychopathology, development
1	mental health, psychiatric disorders	**21**	social skills, social cognition
2	behavioral problems, behavioural problems	**22**	heart rate, respiratory sinus
3	autism spectrum, autism	**23**	borderline personality, personality disorder
4	aggressive behavior, reactive aggression	**24**	functional mri, resonance imaging
5	generalized anxiety, anxiety	**25**	bullying victimization, bullying
6	maternal depression, depression	**26**	bipolar disorder, bipolar
7	suicidal ideation, self harm	**27**	internet addiction, addiction
8	hyperactivity disorder, adhd	**28**	sexual abuse, abuse neglect
9	behavior checklist, behaviour checklist	**29**	parental cancer, cancer
10	eating disorders, bulimia nervosa	**30**	gender differences, gender difference
11	substance abuse, drug abuse	**31**	sleep duration, hyperactivity disorder
12	emotional intelligence, socio emotional	**32**	difficulties questionnaire, learning disabilities
13	traumatic stress, trauma	**33**	adult attachment, attachment
14	depression, maternal depression	**34**	maternal depression, depression
15	generalized anxiety, maternal depression	**35**	refugee children, psychiatric disorders
16	psychological adjustment, socio emotional	**36**	dating violence, violence exposure
17	mindfulness meditation, mindfulness	**37**	congenital heart, heart disease
18	maternal depression, depression	**38**	gender dysphoria, gender identity
19	prefrontal cortex, orbitofrontal cortex	**39**	sexual abuse, maternal depression

The Tables 4, 5, and 6 will further support the outcome of the corresponding networks. *BERTopic* and *Top2Vec* models seem to have more distinguishable topics that can be understood as phenomena. For instance in Table 6, the words 'eating', 'smoking' and 'bulling' are in different topics and in general represent different social phenomena. On the other hand, *LDA* most words are common between topics, such as 'adolescent' and 'study', as depicted from Table 4. Also, in Table 5, the words/bigrams 'depression', 'internet addiction' and 'autism spectrum' represent different topics, while in the same table several words, such as 'abuse' and 'disorder' are common among different topics.

Table 6
The 52 Topics that are produced from BERTopic with the 2 words with the highest scores for each topic

Topic	Words with highest scores	Topic	Words with highest scores
0	problems, behavior	**26**	mindfulness, training
1	mental, health	**27**	reappraisal, strategy
2	autism, autism disorder	**28**	family, parent
3	trauma, ptsd	**29**	self regulation, regulation
4	eating, weight	**30**	empathy, cognitive empathy
5	suicidal, ideation	**31**	smoking, smoker
6	emotion, regulation	**32**	attachment, representation
7	stress, cortisol	**33**	adjustment, student
8	disorder, bipolar	**34**	immigrant, migrant
9	connectivity, region	**35**	mother, infant
10	peer, victimization	**36**	cd, conduct disorder
11	aggression, high	**37**	chd, congenital
12	violence, exposure	**38**	preterm, vlbw
13	substance, cannabis	**39**	alexithymia, difficulty feeling
14	anger, anger regulation	**40**	factor, der
15	bullying, victim	**41**	abuse, abused
16	self, self self	**42**	student, school
17	depressive, symptoms	**43**	romantic, romantic relationship
18	resilience, protective	**44**	student, competency
19	dysregulation, emotional dysregulation	**45**	aggression, parent aggression
20	sleep, daytime	**46**	pubertal, pubertal timing
21	addiction, internet	**47**	treatment, intervention
22	parenting, parent	**48**	sibling, sibling relationship
23	rumination, depressive	**49**	survivors, cancer
24	mother, depressive	**50**	asthma, adolescent asthma
25	sex, risk	**51**	victimization, victimized

Conclusion and Discussion

Text mining systematic review is used as an efficient method for phenomena detection. In this research, three topic modeling algorithms, *LDA*, *Top2Vec* and *BERTopic*, are applied, and the results of the analysis show that *BERTopic* produced more meaningful topics than *Top2Vec* and *LDA*.

The networks, which are produced from *BERTopic*, contain distinguishable topics that can easily be identified as meaningful phenomena that correspond to the main subject of each dataset. Specifically, for abstracts' networks presented in E10 and E12, the names of the topics include mostly different words. In addition, the most common words that were observed during exploratory analysis, such as 'cooperation', 'game', 'group' for the prisoner's dilemma dataset, and 'child', 'study', 'adolescent' for the emotional regulation dataset, as shown in Figures A2b and A4b, are not included in the names of the topics. This can be attributed to the fact that *BERTopic* excludes common words. Moreover, the high topic quality values of abstracts' datasets show the coherence of the topics and the diversity among them. Hence, for *BERTopic* the topic quality results agree with the observations of the networks.

On the other hand, the performance of *LDA* and *Top2Vec* varies between the datasets. Even though *LDA* has the highest topic quality value for keywords of the emotional regulation dataset, it has the lowest value for keywords of the prisoner's dilemma dataset. Also, given the respective networks, as depicted in E1 and E3, only a few phenomena can be identified and their names contain mainly the same most frequent words of each dataset. As for *Top2Vec*, the produced networks and graphs indicate that topics correspond to relevant phenomena. The usage of bigrams enhanced the creation of logical topics, which differ from each other in several words, and at the same time remain relevant to each other due to common words. However, *Top2Vec* topic quality values do not represent the results from the networks since they are significantly lower than the other models, as shown in Tables 1 and 2. The aforementioned findings lead to the conclusion that *BERTopic* is preferred over *LDA* and *Top2Vec* for identifying relevant phenomena in published literature text data. Consequently, the number of phenomena

that are generated by *BERTopic* is considered optimal.

One strength of the present study is that for detecting the relevant phenomena, *Top2Vec* and *BERTopic* are used, which are two recently introduced topic modeling algorithms that use state-of-the-art methods to create the embedding space. Specifically, for *BERTopic*, the embedding space is produced by using a *BERT* model, which creates highly context-specific representations as opposed to other embedding methods, such as *Word2Vec* and *Glove*, which produce the same representations without taking into consideration the context of the surrounding words.

In many topic modeling tasks, topic coherence measures are used to evaluate the performance of the models. Nevertheless, in this thesis it should be highlighted that not only topic coherence but also topic diversity is calculated for evaluating the models' performance in order to produce more objective results. These two metrics are combined in one measure, named topic quality, which is considered more appropriate to underline the relations between the topics/phenomena compared to using one of the two metrics.

One of the limitations of this research is that *BERTopic* and *Top2Vec* contain randomness on their results. Even though I did not observe wide variation in the results while running the models, it is suggested in future work to repeat the execution for a considerably large number of times and calculate the average of the results.

As already mentioned, the networks that were produced from *BERTopic* are very informative regarding the identification of the phenomena, and show clearly the connections between them. However, focusing on the topic quality values of keywords' datasets, it is observed that *LDA* and *Top2Vec* have higher values than *BERTopic*. This might be due to the fact that a sentence encoder is used for this algorithm. Nevertheless, for the keywords' datasets it might not perform as well as in abstracts' datasets because of the absence of context connections since there are no full sentences, but comma-separated words/keywords. It should also be noted that I chose the embedding language model of *"paraphrase-MiniLM-L3-v2"*, considering that it has the best trade-off of performance and speed. However, in future work, it is proposed to use other embedding language models, or even better instead of BERT to use GPT-3.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *International conference on database theory*, 420–434.

Aggarwal, C. C., & Zhai, C. X. (2012). An introduction to text mining. *Mining text data* (pp. 1–10). Springer.

Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*, *27*(4), 509–523. https://doi.org/10.1177/0894439309332293

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Bird, S., Klein, E., & Loper, E. (2008). Nltk documentation. *Online: accessed April*.

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

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The philosophical review*, *97*(3), 303–352.

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766.

Breck, E., Choi, Y., & Cardie, C. (2007). Identifying expressions of opinion in context. *IJCAI*, *7*, 2683–2688.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia conference on knowledge discovery and data mining*, 160–172.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. *Proceedings of human*

language technology conference and conference on empirical methods in natural language processing, 355–362.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, *23*(1), 155–162.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, *96*(34), 226–231.

Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *biometrics*, *21*, 768–769.

Friedman, K. (2003). Theory construction in design research: Criteria: Approaches, and methods [Common Ground]. *Design Studies*, *24*(6), 507–522. `https://doi.org/https://doi.org/10.1016/S0142-694X(03)00039-5`

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Haig, B. D. (2013). Detecting psychological phenomena: Taking bottom-up research seriously. *The American journal of psychology*, *126*(2), 135–153.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. MIT press.

Herfel, W. E. (1995). *Theories and models in scientific processes: Proceedings of afos'94 workshop, august 15-26, mądralin and iuhps'94 conference, august 27-29, warszawa* (Vol. 44). Rodopi.

Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 604–613.

Joachims, T. (1996). *A probabilistic analysis of the rocchio algorithm with tfidf for text categorization.* (tech. rep.). Carnegie-mellon univ pittsburgh pa dept of computer science.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning*, 1188–1196.

Li, D., Wang, Z., Wang, L., Sohn, S., Shen, F., Murad, M. H., & Liu, H. (2016). A text-mining framework for supporting systematic reviews. *American journal of information management*, *1*(1), 1.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93–118.

Markovsky, B., & Webster Jr, M. (2007). Theory construction. *The Blackwell encyclopedia of sociology*.

McInnes, L., & Healy, J. (2017). Accelerated hierarchical density based clustering. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 33–42.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic reviews*, *4*(1), 1–22.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*.

Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International conference on data science and advanced analytics (DSAA)*, 165–174.

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). Octis: Comparing and optimizing topic models is simple! *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263–270.

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research synthesis methods*, *2*(1), 1–14.

Usai, A., Pironti, M., Mital, M., & Mejri, C. A. (2018). Knowledge discovery out of text data: A systematic review via text mining. *Journal of Knowledge Management*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

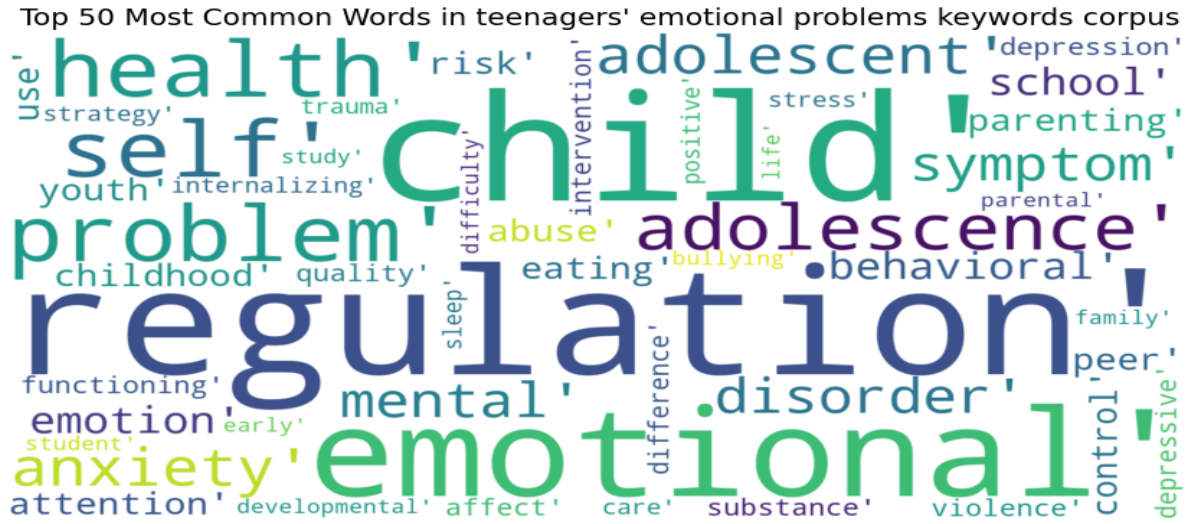
van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133.

van Lissa, C. J. (2021). Mapping phenomena relevant to adolescent emotion regulation: A text-mining systematic review. *Adolescent research review*, 1–13.

Appendix A

Figure A1

(a) *Word Cloud for 50 most common words in teenagers' emotional problems keywords corpus*



(b) *Frequencies plot for 25 most common words in teenagers' emotional problems keywords corpus*

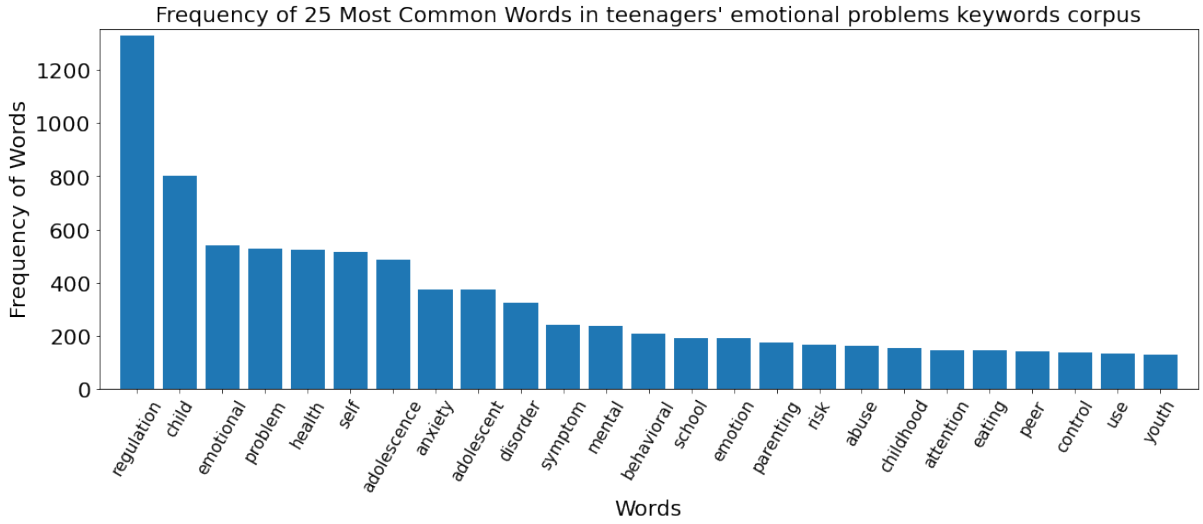
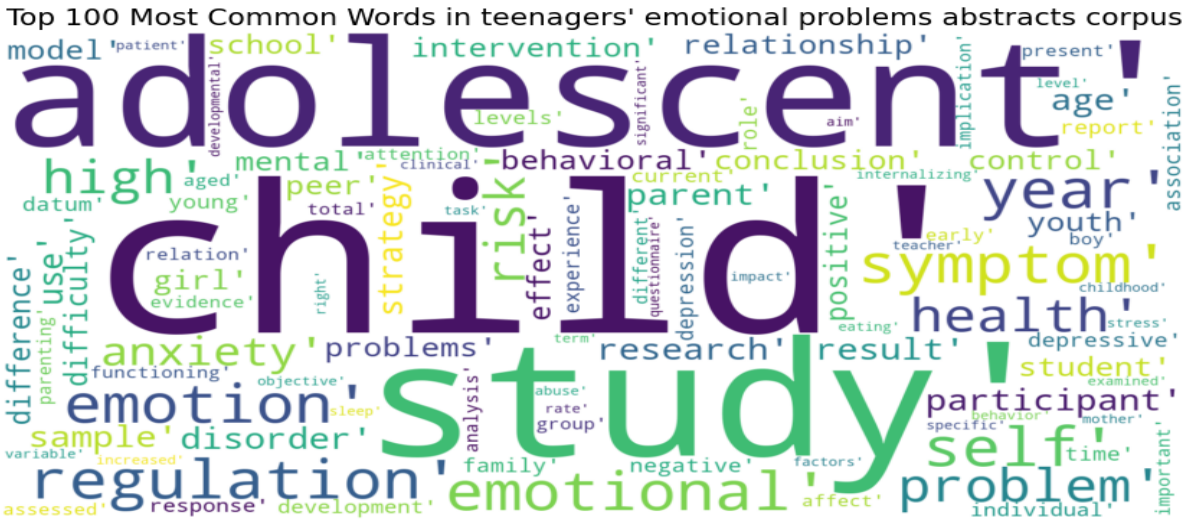


Figure A2

(a) *Word Cloud for 50 most common words in teenagers' emotional problems abstracts corpus*



(b) *Frequencies plot for 25 most common words in teenagers' emotional problems abstracts corpus*

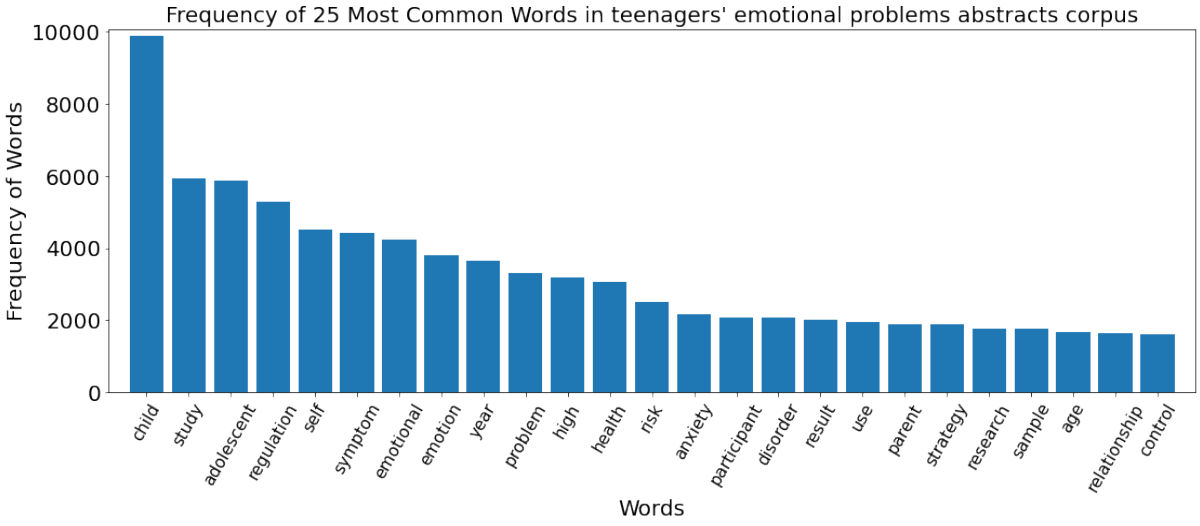
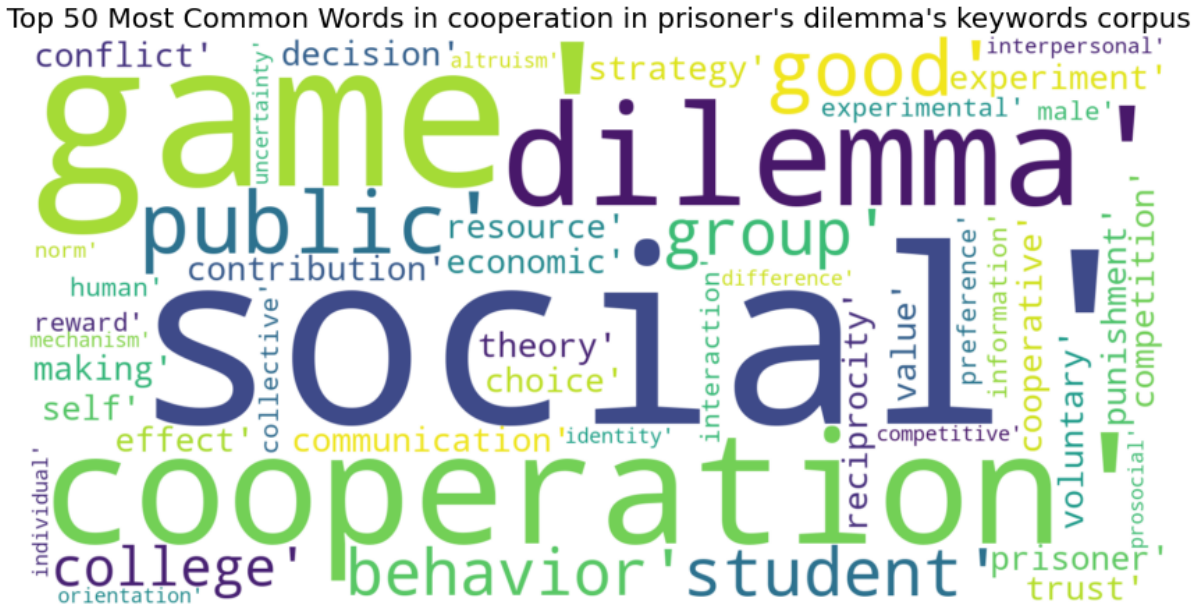


Figure A3

(a) *Word Cloud for 50 most common words in prisoner's dilemma's keywords corpus*



(b) *Frequencies plot for 25 most common words in prisoner's dilemma's keywords corpus*

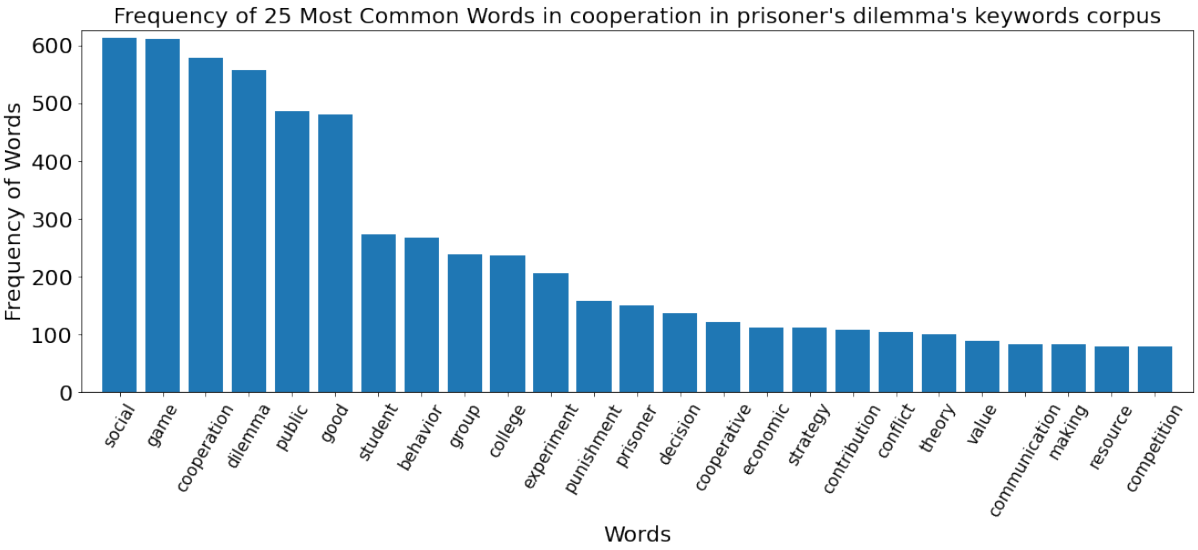
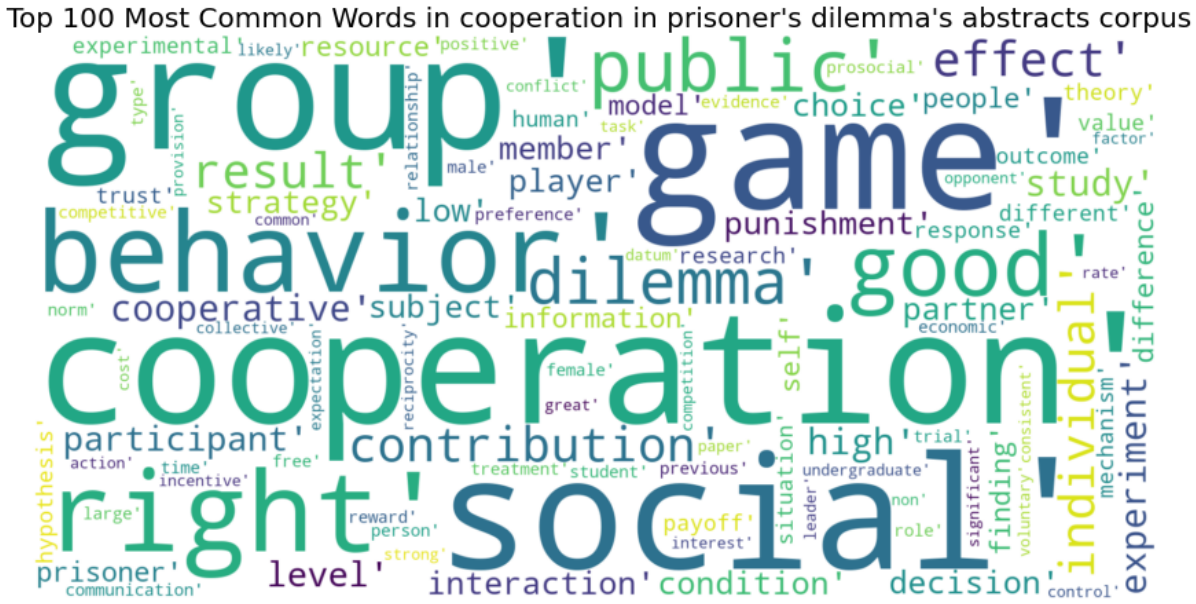


Figure A4



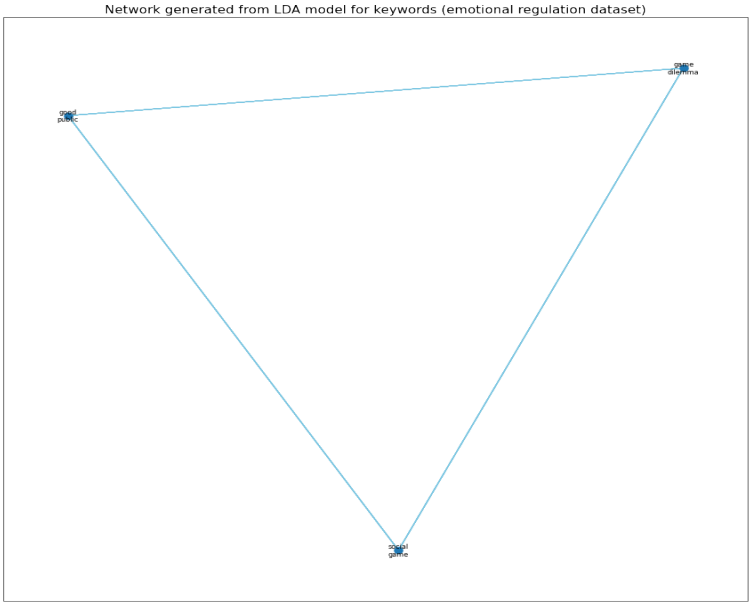
(a) *Word Cloud for 50 most common words in prisoner's dilemma's abstracts corpus*

(b) *Frequencies plot for 25 most common words in prisoner's dilemma's abstracts corpus*

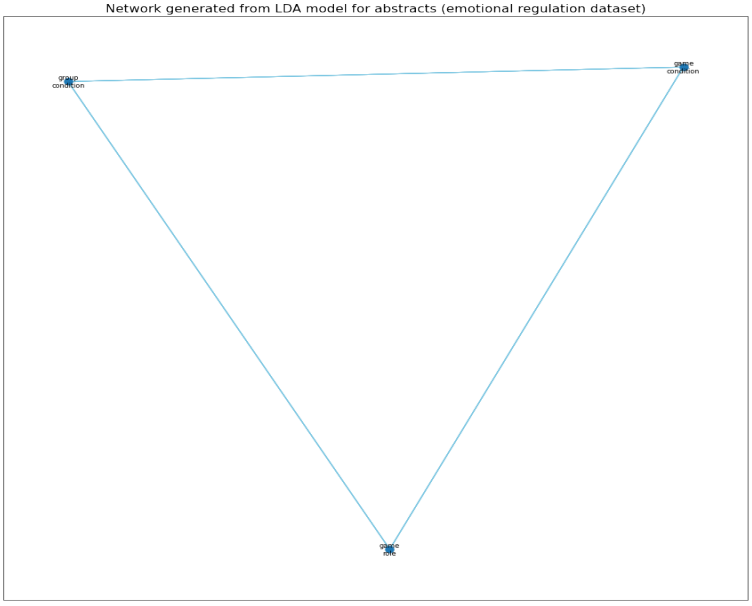
Appendix B

Figure B1

LDA network with 3 topics



(a)

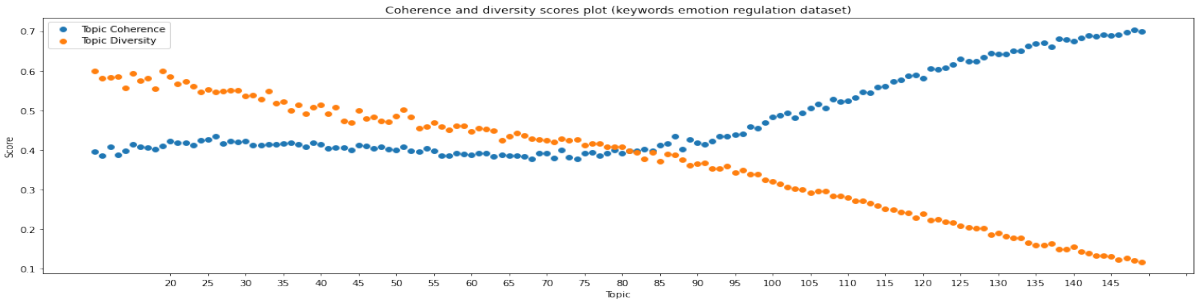


(b)

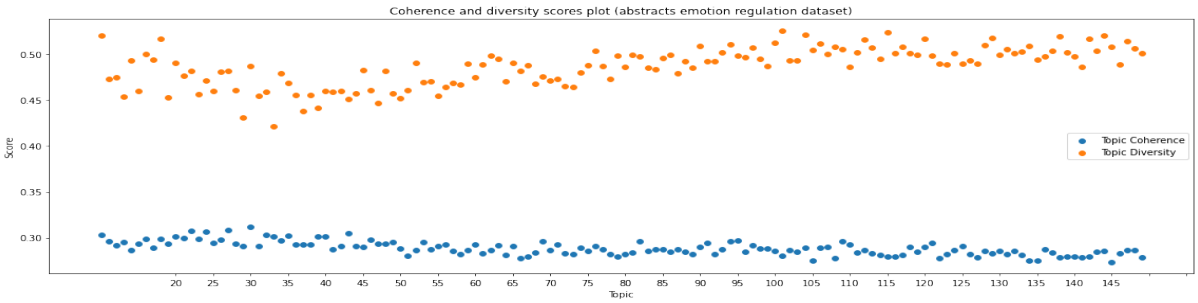
Appendix C

Figure C1

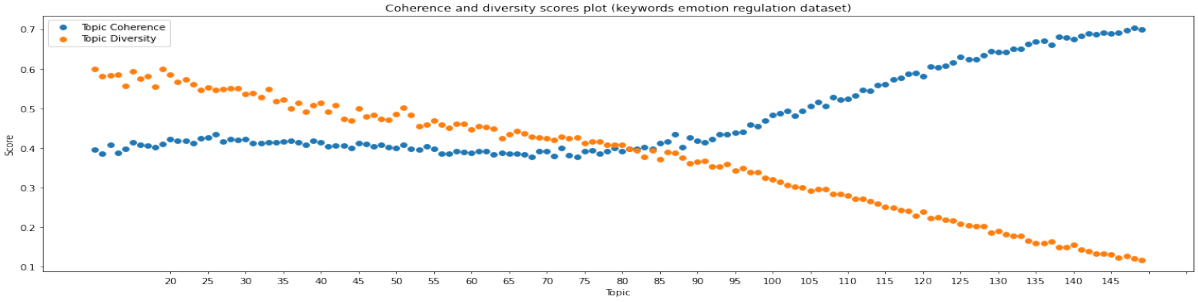
Topic coherence and topic diversity plots for LDA



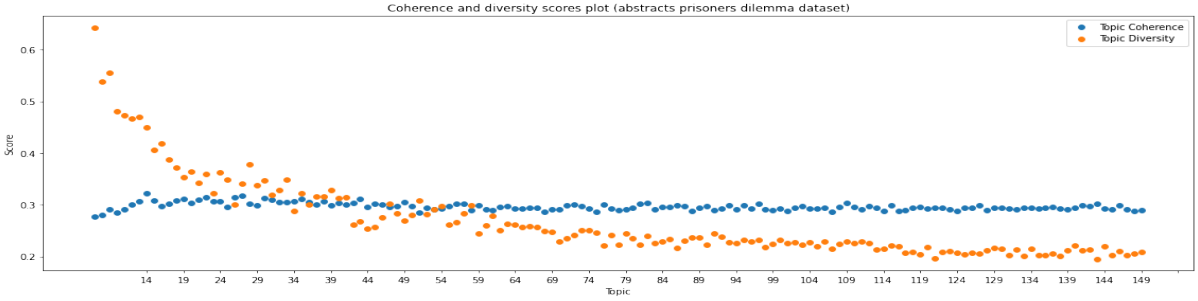
(a)



(b)



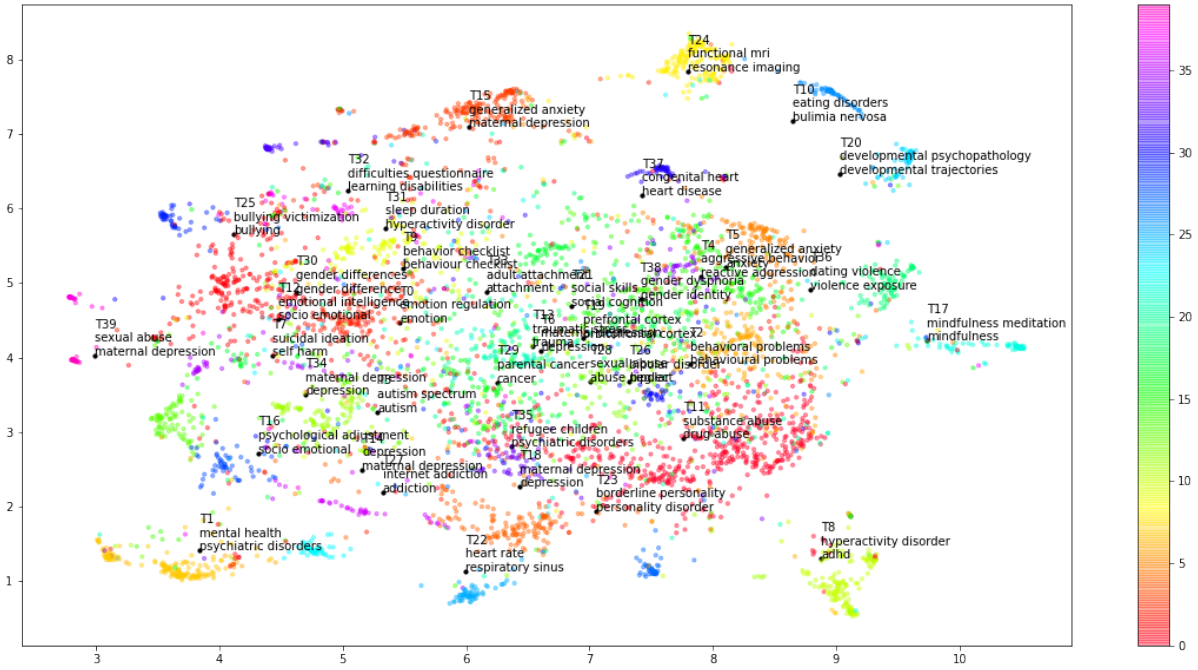
(c)



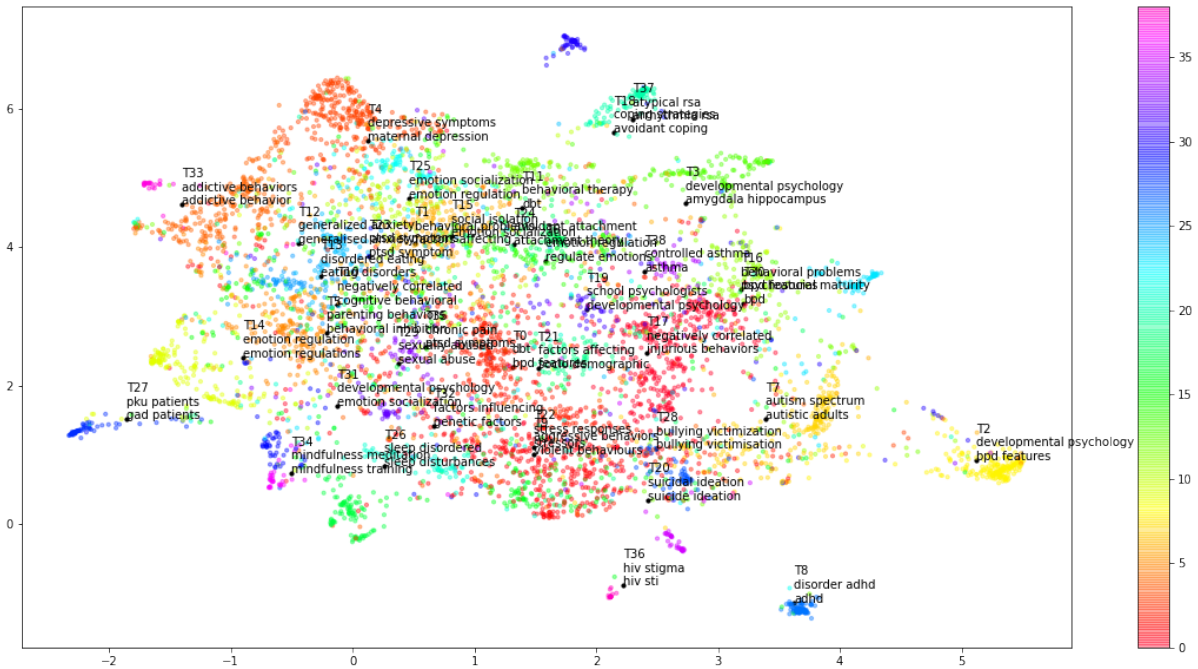
(d)

Figure C2

(a) *Top2Vec word embeddings for 40 topics keywords emotional regulation dataset*



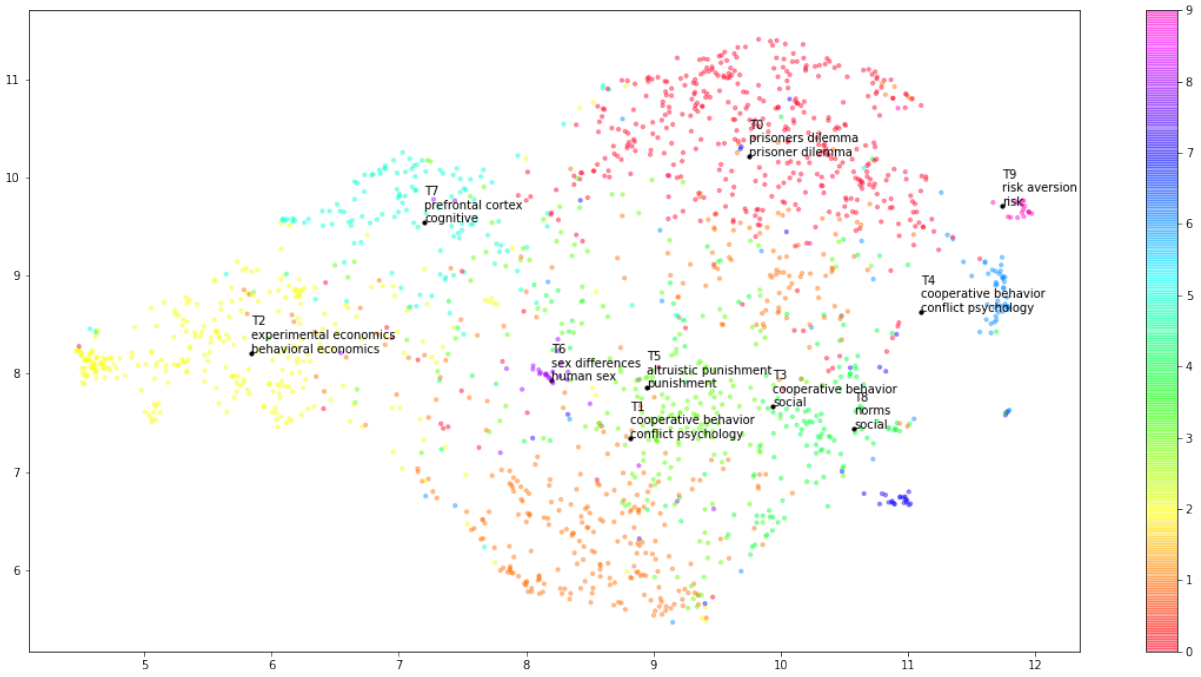
(b) *Top2Vec word embeddings for 39 topics abstracts emotional regulation dataset*



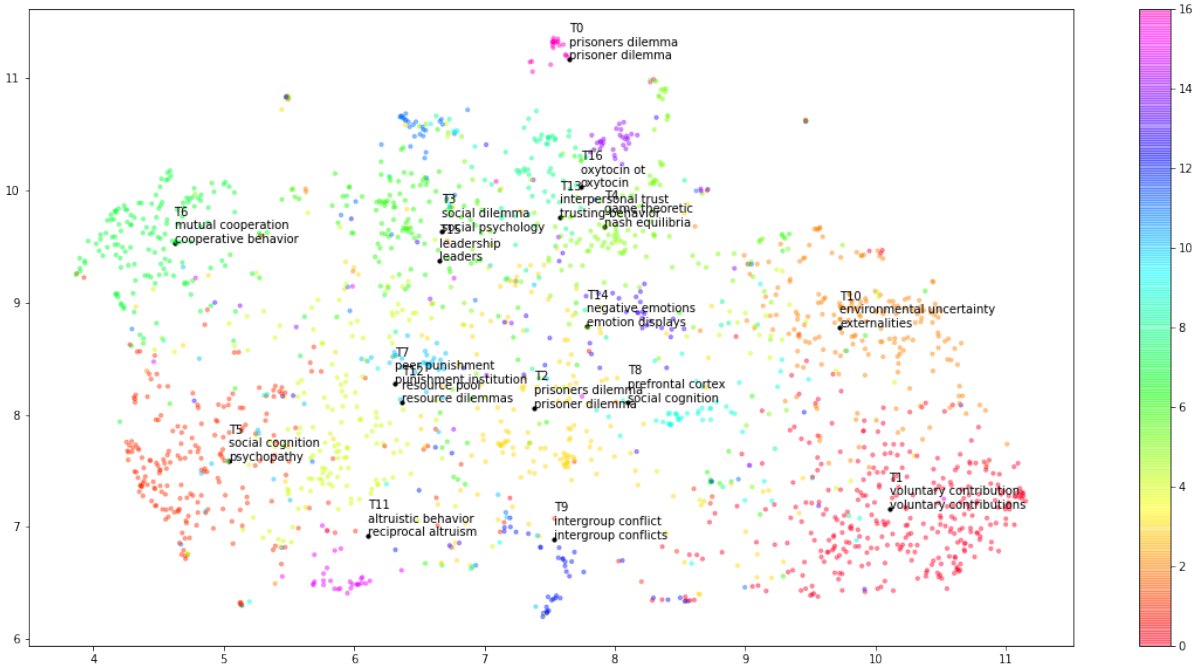
Appendix D

Figure D1

(a) *Top2Vec word embeddings for 10 topics keywords prisoner's dilemma dataset*



(b) *Top2Vec word embeddings for 17 topics abstracts prisoner's dilemma dataset*



Appendix E

Figure E1

Network LDA keywords emotional regulation dataset

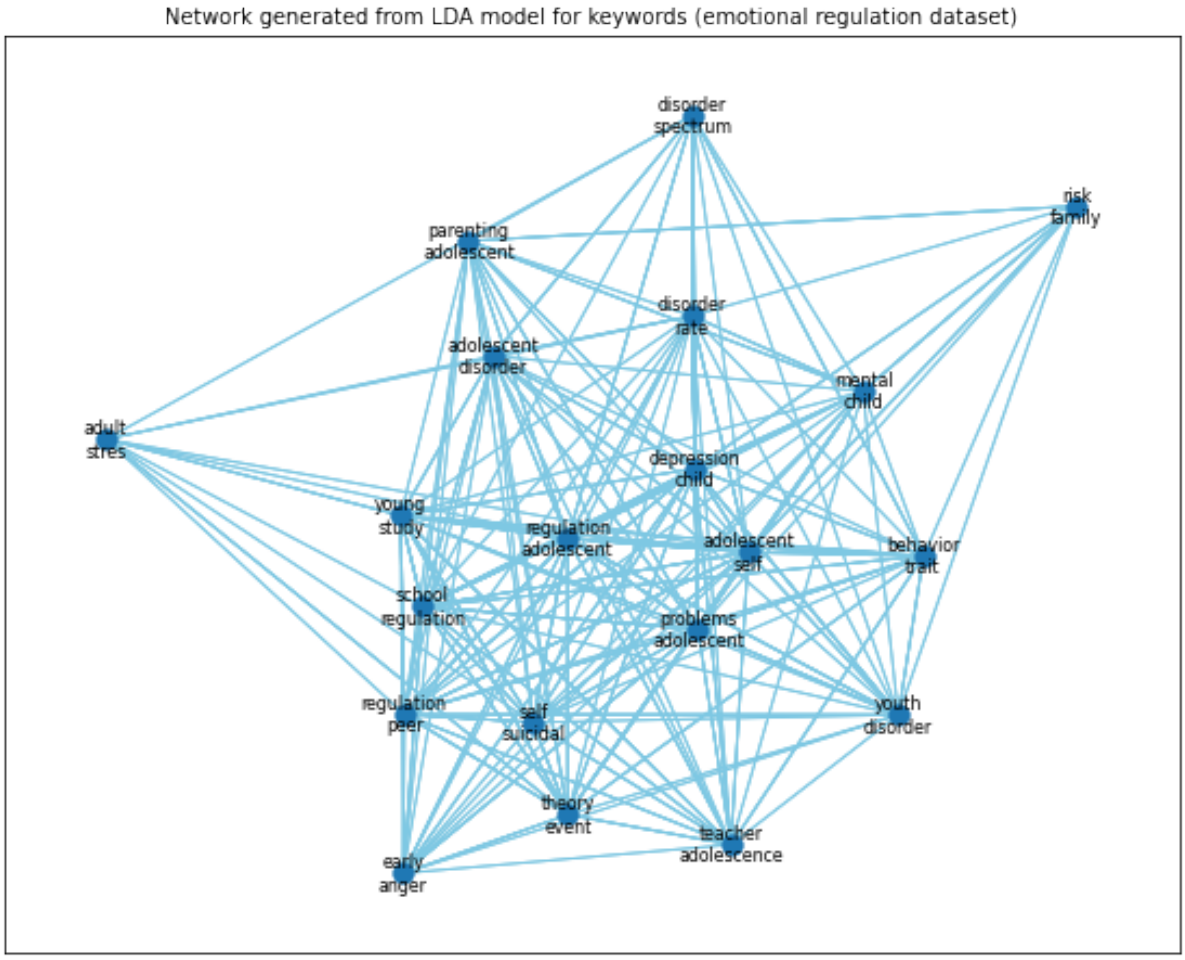


Figure E2

Network LDA abstracts emotional regulation dataset

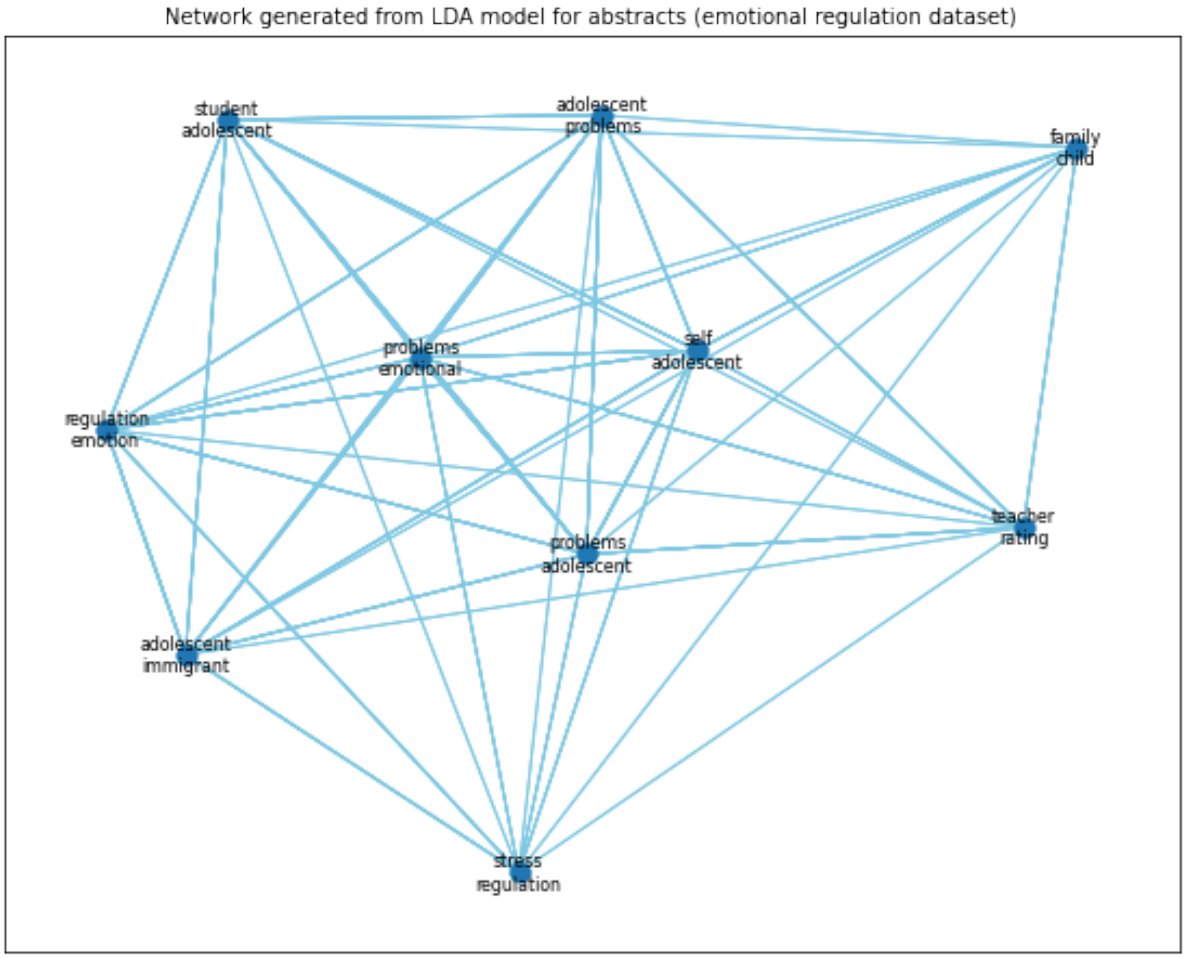


Figure E3

Network LDA keywords prisoner's dilemma dataset

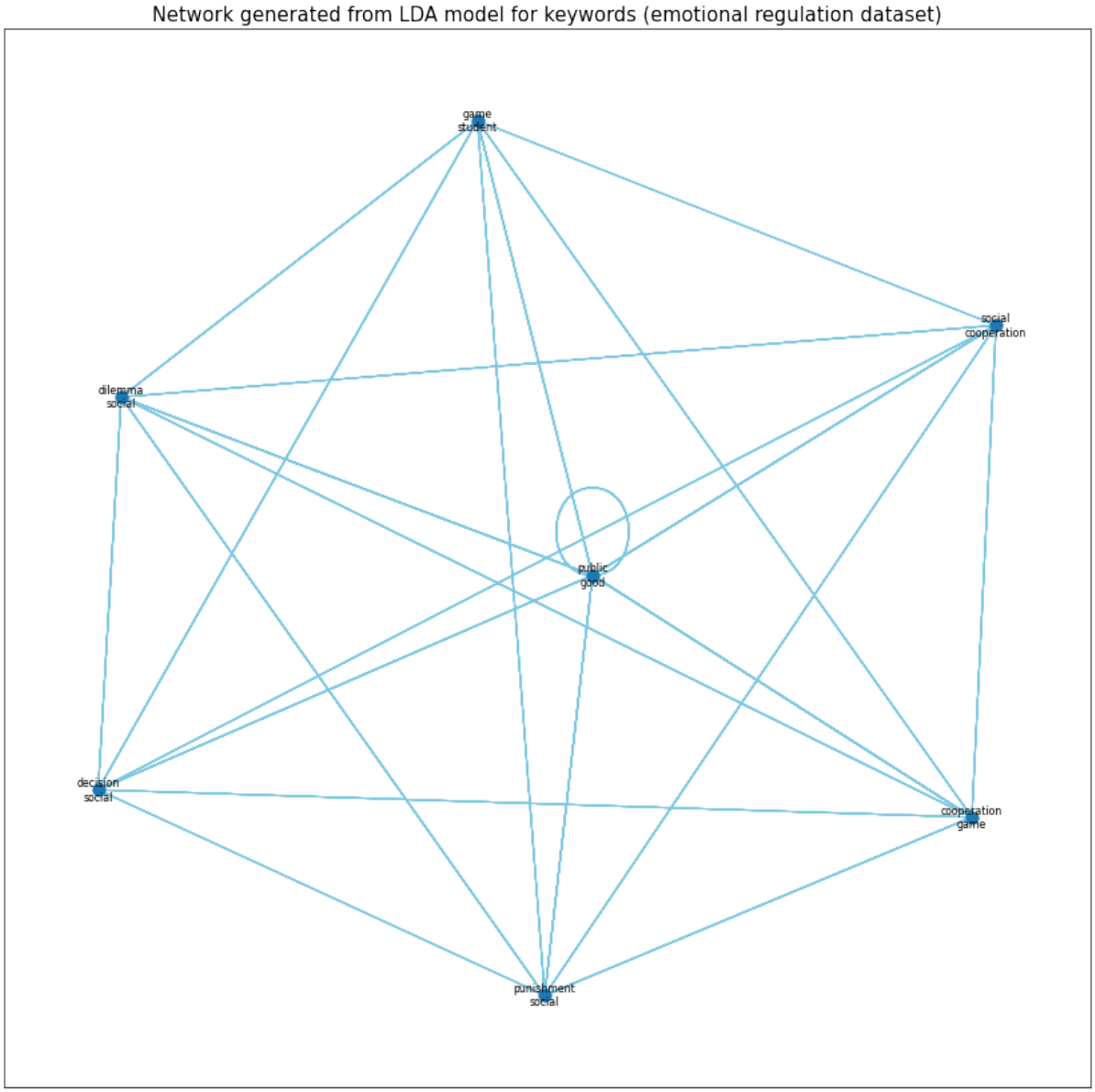


Figure E4

Network LDA abstracts prisoner's dilemma dataset

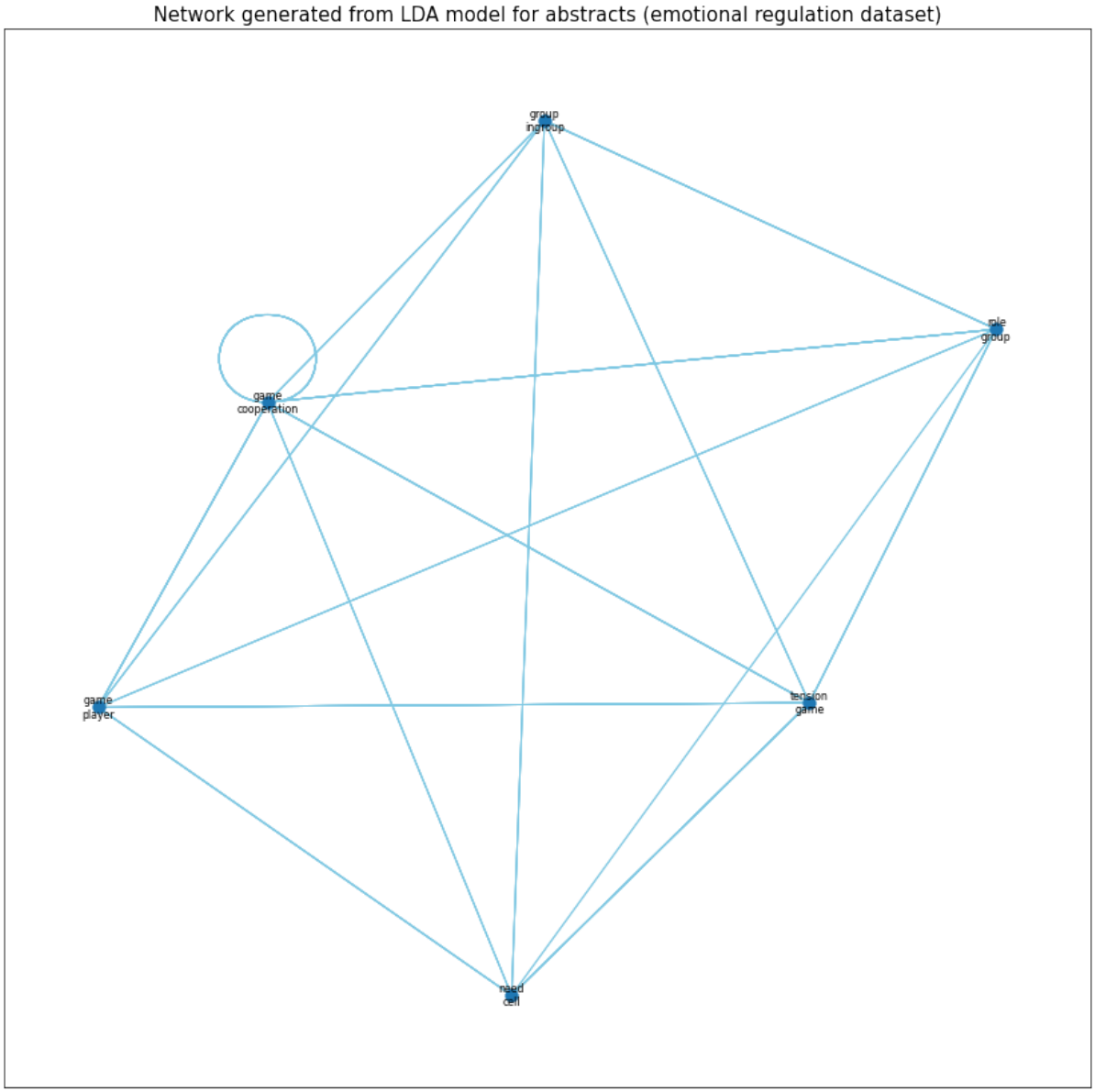


Figure E5

Network Top2Vec keywords emotional regulation dataset

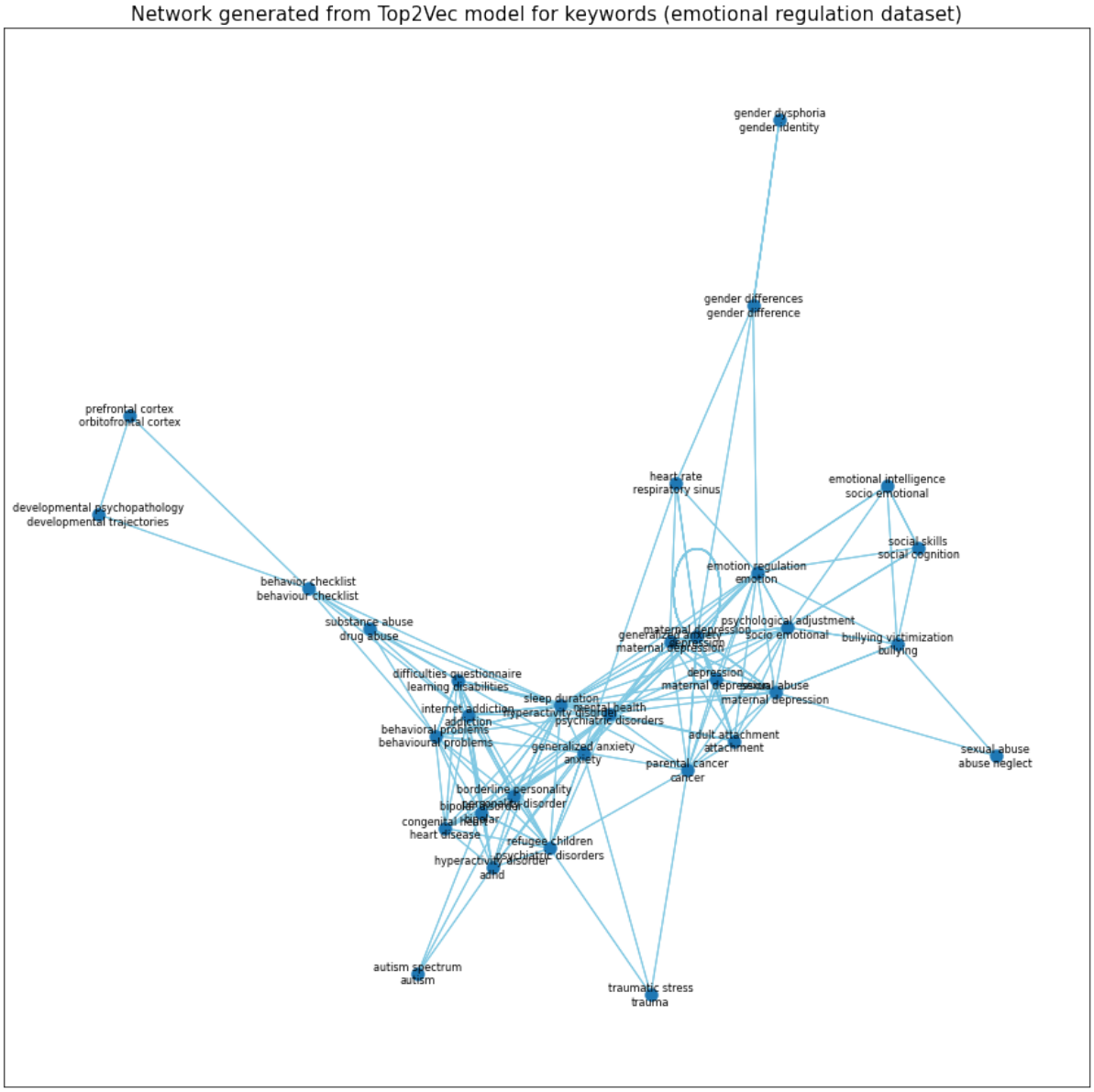


Figure E6

Network Top2Vec abstracts emotional regulation dataset

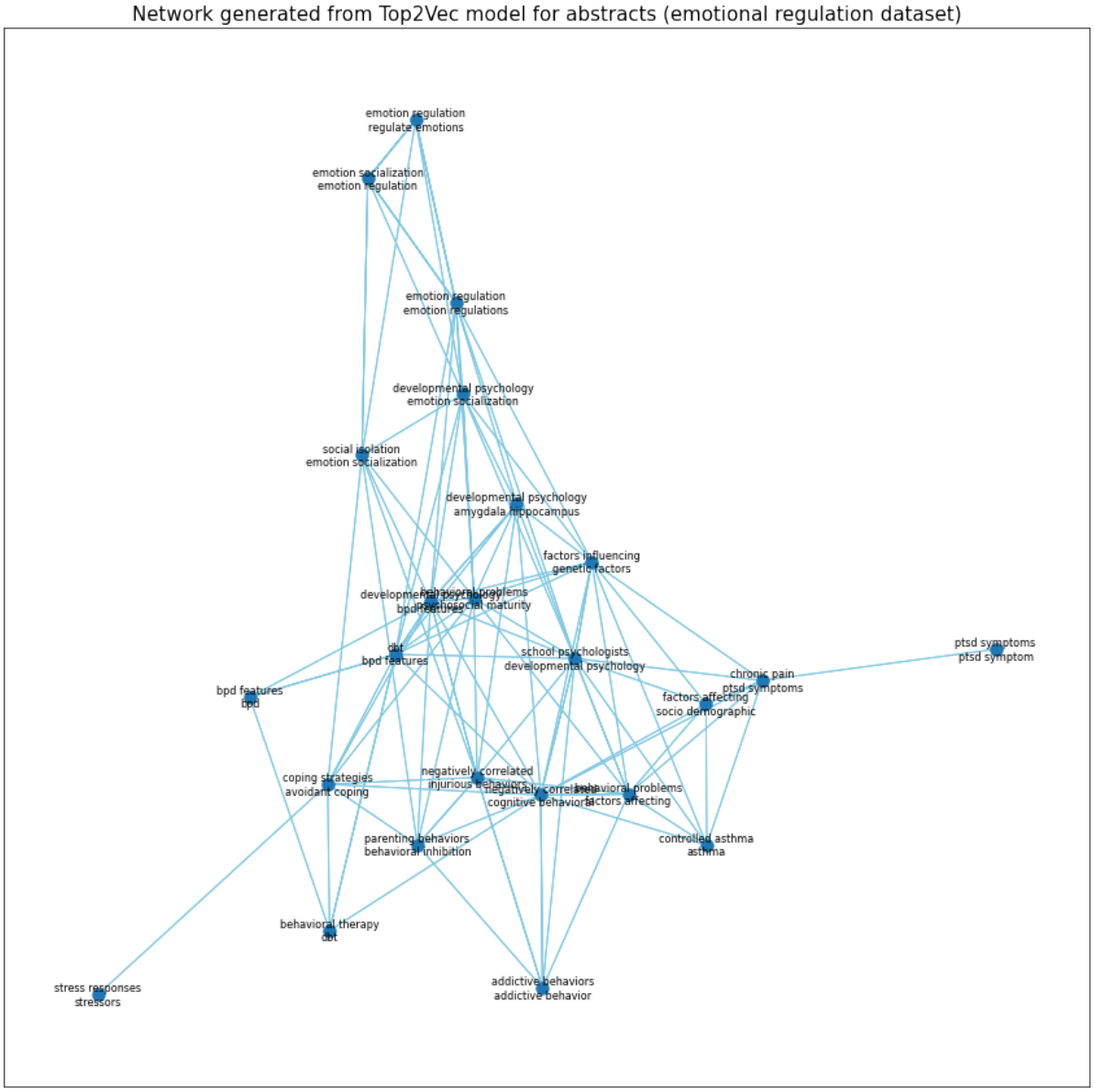


Figure E7

Network Top2Vec keywords prisoner's dilemma dataset

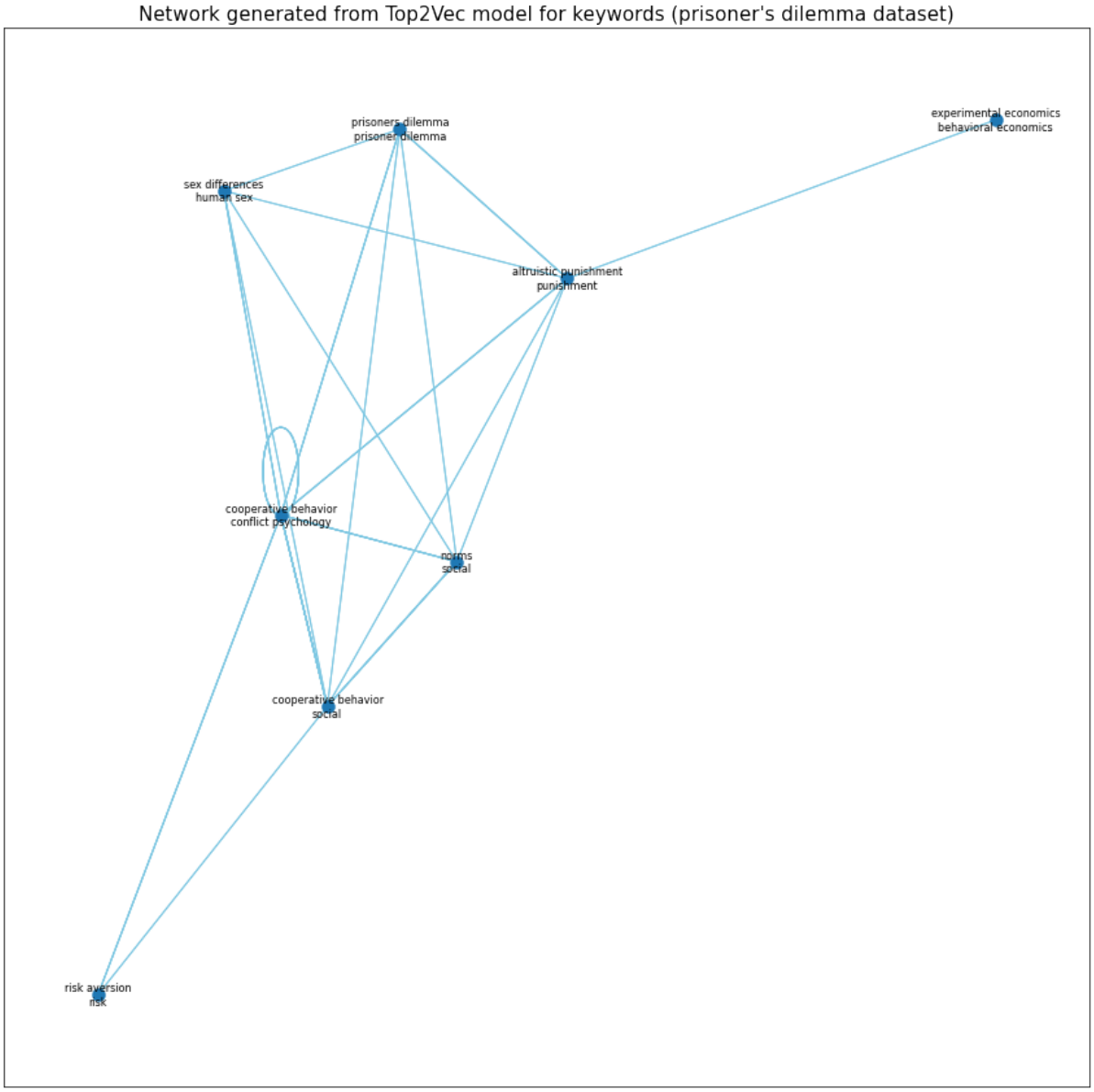


Figure E8

Network Top2Vec abstracts prisoner's dilemma dataset

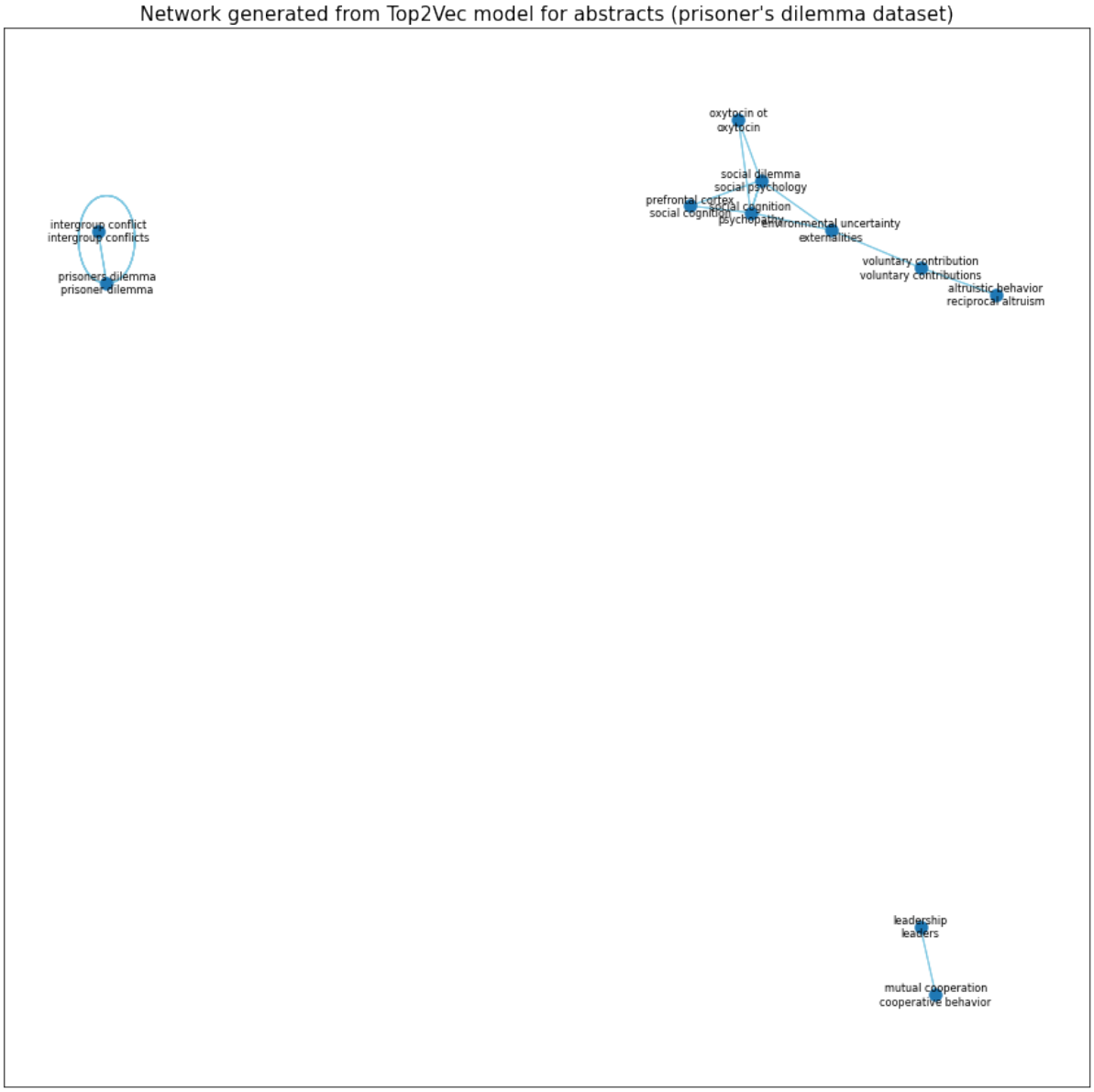


Figure E9

Network BERTopic keywords emotional regulation dataset

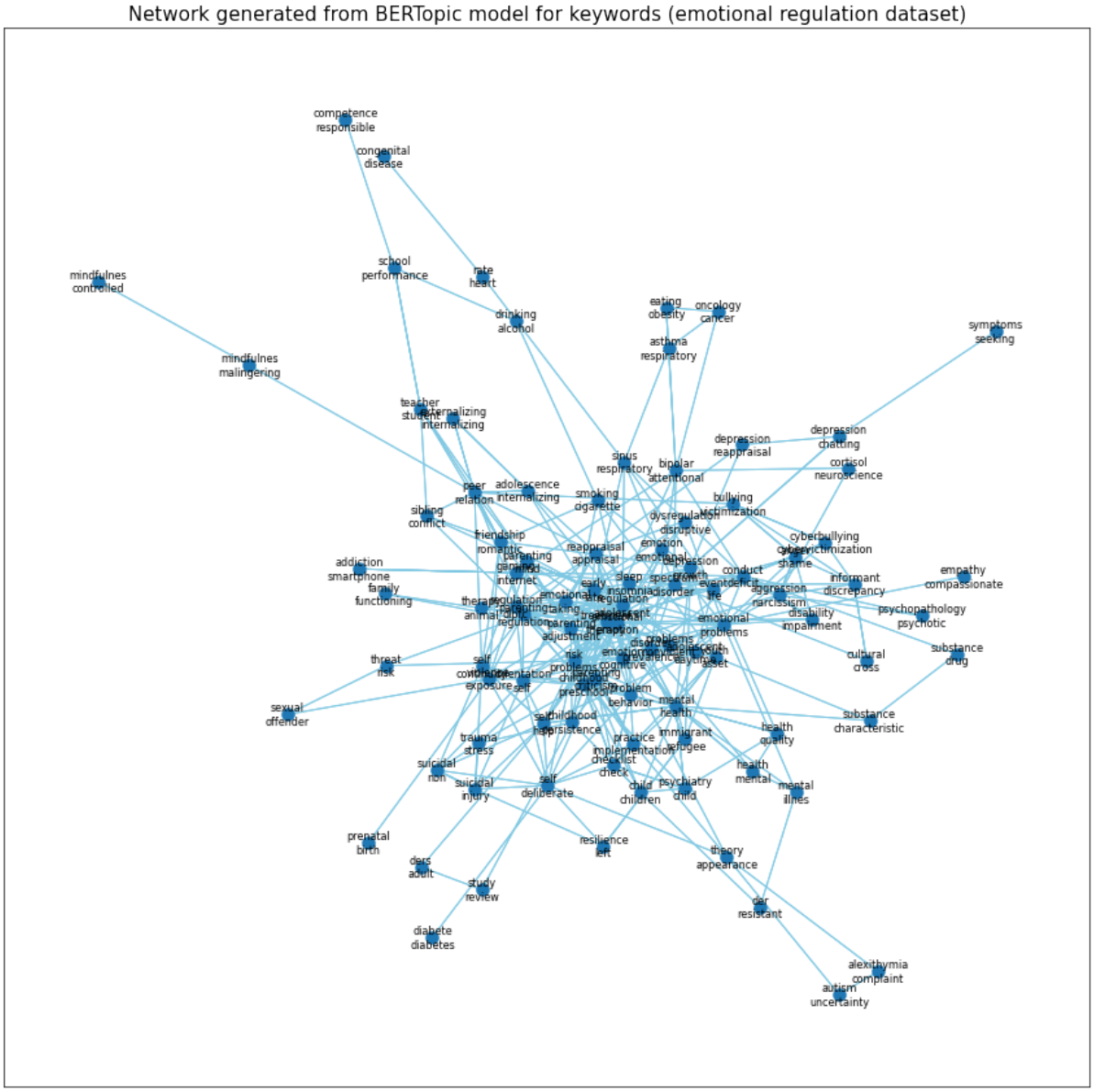


Figure E10

Network BERTopic abstracts emotional regulation dataset

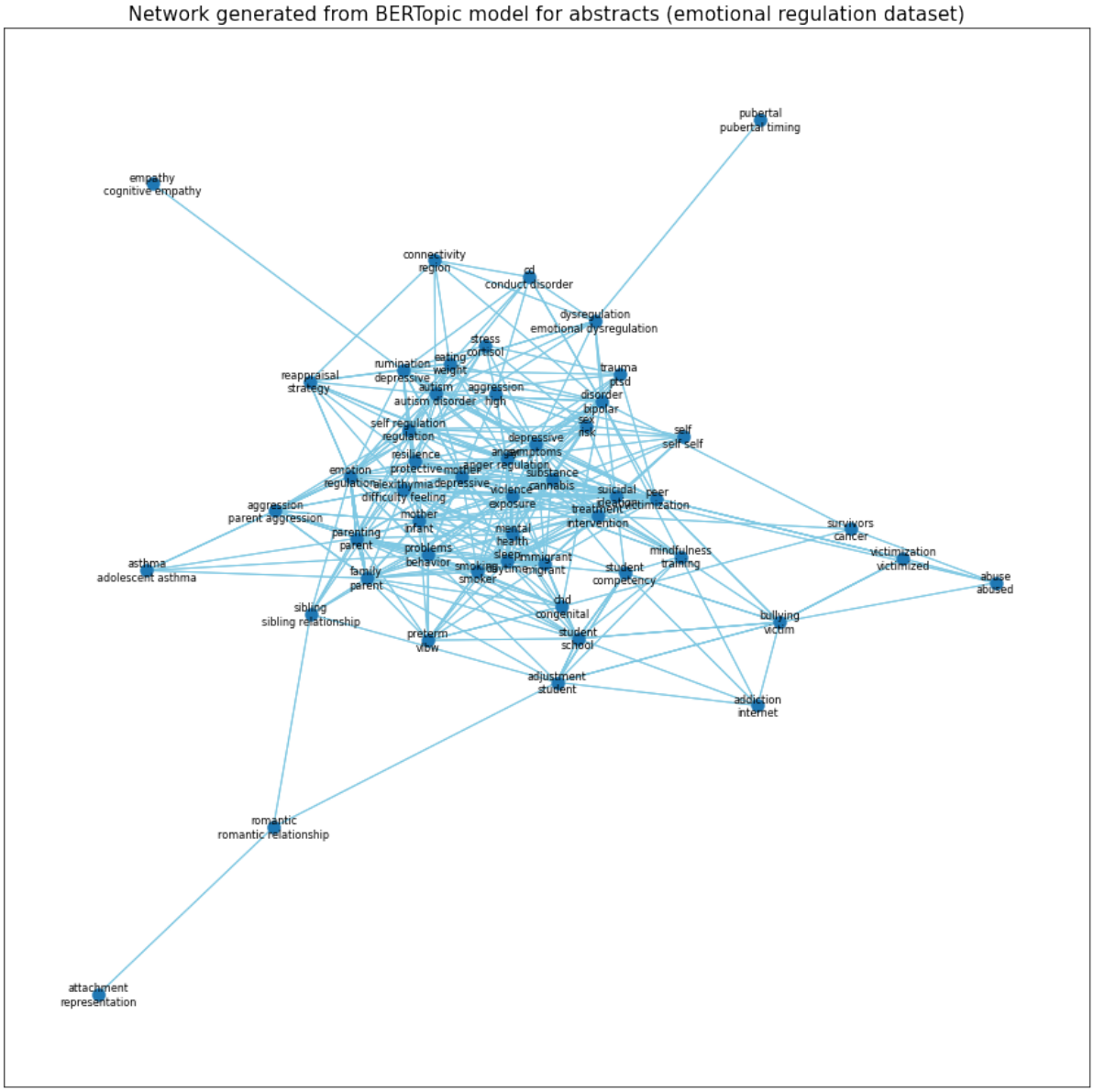
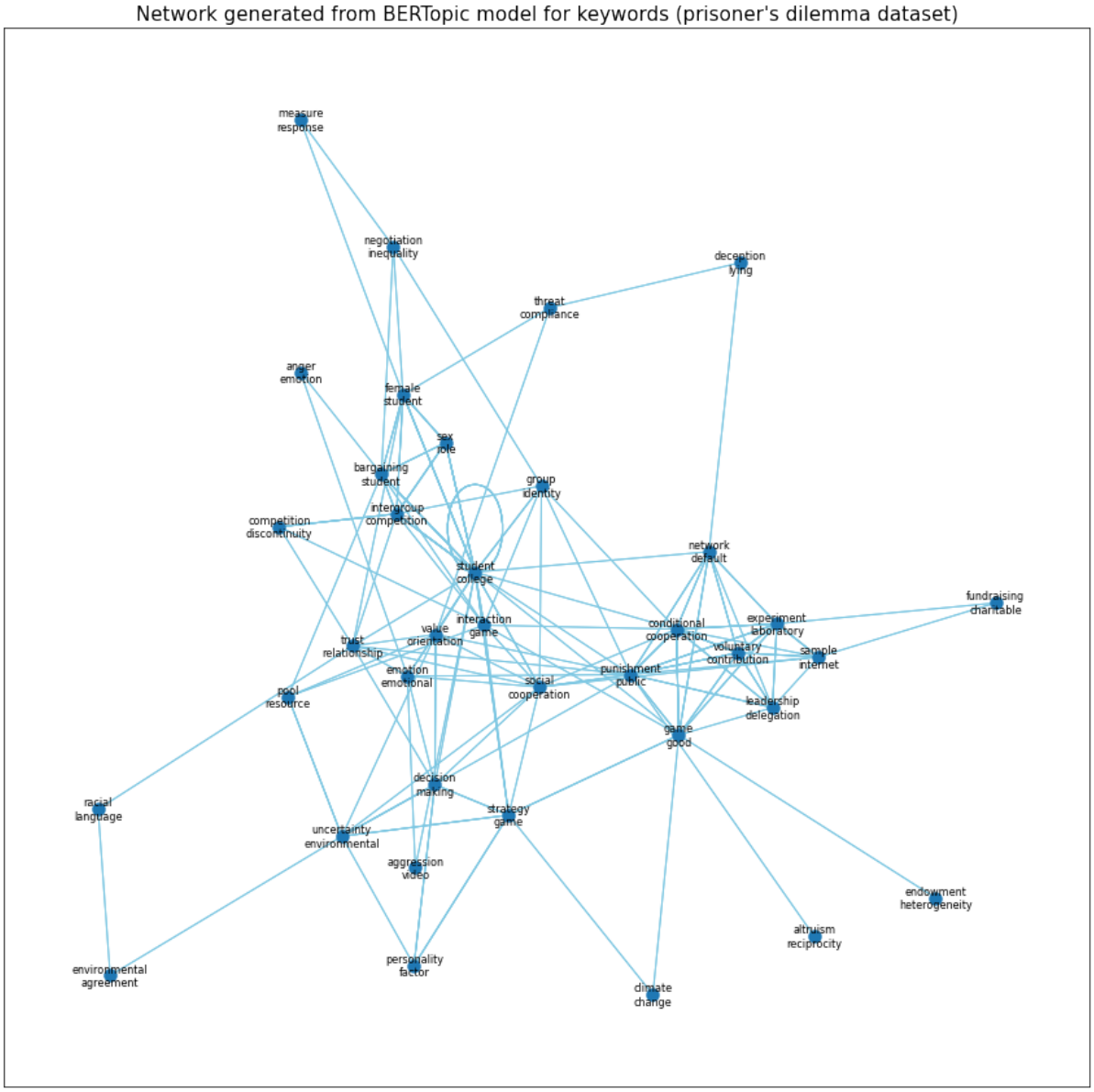


Figure E11

Network BERTopic keywords prisoner's dilemma dataset



B

Figure E12

Network BERTopic abstracts prisoner's dilemma dataset