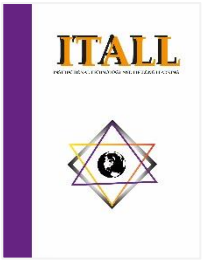
Atıf/ Citation: Akbay, T. (2022). Modeling education studies indexed in Web of Science using natural language processing. *Instructional Technology and Lifelong Learning, 3*(2), 129-143. DOI: 10.52911/itall.1193460



Instructional Technology and Lifelong Learning Vol. 3, Issue 2, 129-143 (2022)

https://dergipark.org.tr/tr/pub/itall

ITALL

Research Article

ISSN: 2717-8307

Modeling Education Studies Indexed in Web of Science Using Natural Language Processing

Tuncer AKBAY*[1]

ARTICLE INFO

Article history:
Received:
Accepted:
Online:
Published:

Keywords:
Topic modeling
Machine Learning
Education
NLP
Artificial intelligence
Top2Vec algorithm

ABSTRACT

Easier access to information and resources allowed researchers to conduct more studies and publish most of them electronically. They are indexed in scholarly citation databases such as Web of Science and Scopus. These databases index huge volumes of research reports. Even though they offer search engine filtering options, it is still hard to locate the publications in which their contents are closely related. Artificial intelligence technologies, such as Natural Language Processing, allow documents to be categorized based on their content. Top2Vec is an unsupervised topic modeling algorithm that enables users to categorize documents semantically. The purpose of the current study is twofold: (1) to provide users with the ability to group documents applying Natural Language Processing techniques, and (2) to reveal the topics with the highest number of articles indexed in the 'education scientific disciplines' category within the Web of Science Core Collection scholarly database in 2021. Colab notebook used to type Python codes for executing Top2Vec algorithm. This study yielded 68 distinct topics among the 8125 articles published in 2021 and indexed in the Web of Science database under the Education Scientific Disciplines category. After modeled topics were ranked from the topic having the largest number of documents (i.e., N=549) to the topic having the least number of documents (i.e., N=29), the first eight topics' findings were presented and discussed. These eight most studies topics are listed as follows: Physics (N=549), online education and covid (N=438), Chemistry (N=381), Math and Reasoning (N=377), Psychology and Emotions (N=257), Educational Diversity (N=228), Health and Life (N=223), Mentoring and Leadership (N=204).

* *Corresponding Author, tuncerakbay@mehmetakif.edu.tr*
[1]*Burdur Mehmet Akif Ersoy University, Türkiye*

Web of Science Atıf İndeksinde Yer Alan Eğitim Araştırmalarının Doğal Dil İşleme Yöntemiyle Modellenmesi

MAKALE BİLGİ

Makale Geçmişi:
Geliş: 23/10/2022
Kabul: 06/12/2022
Çevrimiçi: 09/12/2022
Yayın: 31/12/2022

Anahtar Kelimeler:
Konu modelleme
Makine öğrenmesi
Eğitim Araştırmaları
Doğal Dil İşleme
Top2Vec algoritma

ÖZET

Gelişen teknolojiyle birlikte bilgi kaynaklarına erişim daha kolay hale geldi. Bu durum araştırmacıların kısa sürede daha fazla yayın yapmasına ve büyük birçoğunun elektronik olarak yayınlanmasına ve depolanmasına olanak sağladı. Akademik yayınların büyük bir kısmı Web of Science ve Scopus gibi bilimsel veri tabanlarında indekslenirler ve ilgili veri tabanlarından erişilirler. Bu veri tabanları binlerce hatta milyonlarca araştırma raporlarını depolar. Web of Science ve Scopus gibi indexler abonelik tabanlı erişim sağladıkları veri tabanlarından veri almak için arama motoru ve filtreleme seçenekleri sunsalar da içeriklerinin yakından ilişkili olduğu yayınları bulmak yine de zordur. Doğal dil işleme gibi yapay zekâ teknolojileri, belgelerin içeriklerine göre kategorilere ayrılmasını sağlar. Top2Vec, kullanıcıların dokümanları anlamsal olarak kategorize etmelerini sağlayan denetimsiz konu modelleme algoritmalarından biridir. Bu çalışmanın amacı iki yönlüdür: (1) Araştırmacılara doğal dil işleme tekniklerini uygulayarak içerikleri gruplama becerisi kazandırmak ve (2) 2021 yılında yayınlanmış olan ve Web of Science'da 'Education Scientific Dsciplines' (Eğitim Bilimsel Disiplinleri) kategorisinde indekslenen makalelerin içeriklerini gruplandırarak en çok yayın yapılan konuları tespit etmektir. Top2Vec algoritmasını çalıştırmak için yazılacak olan Pyhton kodları Google Colab Notebook kullanılmıştır. Bu çalışmada 2021 yılında yayınlanan ve Web of Science veri tabanında Eğitim Bilimsel Disiplinleri kategorisi altında indekslenen 8125 makale arasından 68 farklı konu tespit edilerek her bir konudaki makale sayıları ortaya konulmuştur. Modellenen konular en fazla yayın yapılmış (örn, makale) konudan (N=549) en az yayın yapılmış konuya (N=29) doğru sıralandıktan sonra ilk sekiz konunun içerdiği anahtar kelimeler raporlanmış ve tartışılmıştır. En çok araştırma yapılan bu sekiz konu şu şekilde listelenmiştir: Fizik eğitimi (N=549), Çevrimiçi Eğitim ve Kovid-19 (N=438), Kimya Eğitimi (N=381), Matematik Eğitimi ve Akıl Yürütme (N=377), Psikoloji ve Duygu Durumu (N=257), Eğitimde Kültürel Çeşitlilik (N=228), Sağlık ve Yaşam (N=223), Mentorluk ve Liderlik (N=204).

1. Introduction

Developments in Web technologies have made significant improvement on creating and sharing content in any field. Due to internet and digitalization, it has never been easier to reach information as it is today. Because of easy access to information, researching has speeded up. Increasingly academic papers and research reports are being published day by day. Therefore, it is getting harder to follow the publications regarding your own study/ research area. Indeed, academic databases like Web of Science Core Collection and Scopus provides search engine and some filtering options for readers to locate the most relevant publications. Traditional search engines also offer some refinements and semantic search; however, search results mostly lead searchers to documents location without providing extracted necessary data (Linguamatics, 2022). Such search engines return huge volume of data (i.e., publications, reports, posts) because of single query. As the data get bigger it would become harder to manage. In such cases, we may need an assistance to review the publications or any other texts and extract the valuable information for us. Artificial intelligence technologies enabled such assistant through machine learning and deep learning techniques.

Developments in artificial intelligence allowed people to create systems that manage huge volume of data for offering the best possible solutions in diverse study areas such as health (Sevli, 2019) and education (Anuradha & Velmurugan, 2015). The concepts of data mining refer to such systems that "functions as the machine-driven or convenient extraction of pattern representing knowledge implicitly keep or captured in huge databases, warehouses, the Web, data repositories, and information streams" (Mythili & Mohamed Shanavas, 2014, p.63). It is a promising and developing discipline discovering meaningful hidden patterns from excessive and messy data via application of wide range of techniques, methods, and tools (Anuradha & Velmurugan, 2015; Shmueli, Patel, & Bruce, 2007). Data mining tools are useful for analyzing data comes from any field of work/ study. Therefore, data mining research are conducted in any area of study. Another term worth to mention here is machine learning which is also used in research or problem-solving processes involving any field of science. Machine learning is defined as implicitly (i.e., without programming it explicitly) giving machines (i.e., computers) ability to learn thing to handle the data more efficiently when we human mind remain incapable of doing so after reviewing the data (Mahesh, 2020).

Since machine (i.e., computer) learns from data, machine learning area has wide range of machine learning algorithms repertory. These algorithms are differentiated three main categories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Selection of algorithms to apply machine learning involve

careful consideration. It mostly depends on three criteria: the nature of the problem to be solved, the number of variables, and the model suits the best (Mahesh, 2020). For instance, Supervised Learning algorithms requires the dataset containing both input and output data whereas Unsupervised Learning algorithms runs on the dataset containing only input variables. Additionally, Reinforcement Learning requires an environment wherein the agent operates in. Recently, there is two key areas shine out in artificial intelligence: Natural Language Processing (NLP) and Intelligent Agent (Chang, Yu, Chang, & Yu, 2021; Hirschberg & Manning, 2015).

Natural language can be defined as the collection of words and grammatical rules used (i.e., spoken or written) by humans to communicate each other (Chang, Yu, Chang, & Yu, 2021). Natural language processing allows the languages spoken by human beings to be interpreted by machines (Sevli & Kemaloglu, 2021). NLP is applied by text mining, one of the well-known artificial intelligence technologies, to transform unstructured text in databases or individual documents structured form for analysis to drive conclusion (Linguamatics, 2022). Due to exponential production of texts, thanks to rapid development of information and networks, clustering and classifying huge volume of text and topics of documents without relying on human resources (i.e., domain specialist) became evident (Chang, Yu, Chang, & Yu, 2021). Because of saving time and human resource, it would be wise to use topic modeling, which is one of the natural language processing methods to identify hidden topics within the documents (Karas, Qu, Xu, & Zhu, 2022).

Data scientists uses topic modeling to sort and cluster into topics a large collection of text documents otherwise they cannot be read and sorted by the effort of a person (Angelov, 2020). Ability of topic model is to reveal semantic structure called topic from the vast number of documents including huge volume of texts in it. Topics model may be used for clustering similar documents (Angelov, 2020) which provide people with an opportunity to reach the documents written about similar issues. The use of automatic document grouping technology relying on topic modeling is quite important in terms of speed and effectiveness of information management (Chang, Yu, Chang, & Yu, 2021). As it is mentioned earlier, some databases (i.e., Web of Science, Scopus) offer services like search engines to look for matching keywords and filtering options such as range of publication data or the category of document topic and research field. However, as Angelov (2020) argued topics can overlap each other and they can be subdivided into numerous sub-topics. It may be wrong to rely on predetermined categories of topics, as the most databases generally offer, since those categorization criteria may not fulfil the demands and expectations of us. For instance, some scholarly databases ask authors to choose a category for their manuscripts contains information from diverse disciplines (i.e., STEM related works) among predetermined categories (i.e.,

education, mathematics, engineering) that the topic of their manuscripts suits the best. In such cases, authors must opt one category (let's say mathematics) for their manuscript which possess information regarding more than one category. Thus, those filters deprive searchers, which filtered either education or engineering category, from that manuscript since it has been fallen into education category due to choose of the author.

The purpose of the current study is twofold: (1) provide users with the ability to group documents applying natural language processing techniques, specifically Top2Vec algorithm, and (2) to reveal the topics with the highest number of articles indexed in the 'education scientific disciplines' category within the Web of Science Core Collection scholarly database in 2021.

1.1. Related Studies

Text mining techniques like topic modeling has been used to derive meaningful information from unstructured data. Among those studies, trends in e-learning and distance education research are investigated by Hung (2012), and Zawacki-Richter and Naidu (2016) wherein the datasets contained 689 and 515 publications respectively. All the papers have been published in Computers and Education journal, which is one of the prestigious journals in instructional technology field, for four decade (1976 through 2016) are clustered into four stages by Zawacki-Richter and Latchem () using text mining tools. Moreover, Bohr and Dunlap (2018) analyzed publications between the years of 1990 and 2014 through topic modeling technique to reveal key themes and trends during those years in the field of environmental sociology. Chang, Yu, Chang, and Yu (2021) conducted similar topic modeling research along with co-word analysis in the field of environmental education.

2. Method

2.1. Research Model

The current study applies topic modeling method to discover hidden topics (common themes) from 8125 articles published during the year of 2021 and indexed in Web of Science citation database. Topic Modeling is a family of methods that are powerful smart techniques for facilitating the process of exploratory analysis over huge volume of text collections to extract common themes (Chen, Yu, Zhang, & Yu, 2016; Jelodar et al., 2019). Those methods are widely used in Natural Language Processing.

2.2. Instruments and Tools

Web of Science citation database is used to gather dataset analyzed in this research. For analysis of data, Google Colaboratory, Google Colab in short, notebook used as the integrated development environment. Used

programming language was Python and applied algorithm was Top2Vec. Google Colab is "a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. With Google Colab, it is possible to write and execute code, save, and share our analyses, and access powerful computing resources, all for free from the browser" (Gunawan et al., 2020, p. 2468).

Topic modeling algorithm used for this research is Top2Vec. It is an unsupervised topic modeling (i.e., clustering documents to topics) (Eykens, Guns, & Vanderstraeten, 2022) which means that it does not require any preset number of clusters. The logic behind this algorithm is described as follows: (1) it takes input texts and converts each of them into a vector in semantic space, (2) once the documents embedded into vectoral space, it finds dense cluster of documents through computing the distance between vectors, (3) identify the words pulled those documents together (Angelov, 2020; Eykens, Guns, & Vanderstraeten; Karas, Qu, Xu, & Zhu, 2022).

Top2Vec topic modeling algorithm is presented by Angelov in 2020. It is a relatively new algorithm for topic modeling compared to counterparts such as Latent Dirichlet Allocation and Latent semantic Analysis. Egger and Yu (2022) listed the advantages of Top2Vec topic modeling as follows. Top2Vec

> *"Supports hierarchical topic reduction, allows for multilingual analysis, automatically finds the number of topics, creates jointly embedded word, document, and topic vectors contains built-in search functions (easy to go from topic to documents, search topics, etc.), can work on very large dataset sizes, uses embeddings, so no preprocessing of the original data is needed"* (Egger & Yu, 2022, p. 13).

2.3. Data Analysis and Procedure

First, the *pandas* library was imported as *pd* using the following code.

```
import pandas as pd
```

Next, json file was loaded into pandas data frame and displayed using the following code.

Json file, which contains my dataset, was named as education2021.

```
df = pd.read_json("education2021.json")
df
```

Then, following code was applied to create a list named as *docs* and copy all abstracts to the list. The feature name in the json file should be abstract. Notice that the default name of the column containing the publication abstracts is 'Abstract' in the xlsx document downloaded from Web of Science database.

```
docs = df.Abstract.tolist()
```

Then, Top2Vec was installed through the following code.

```
!pip install Top2Vec
```

Next, the following code is used for applying Top2Vec algorithm and training the dataset.

```
from top2vec import Top2Vec
model = Top2Vec(docs)
```

Because following error was occured: `__init__() got an unexpected keyword argument 'cachedir'`

It was corrected using the following code before the code above.

```
!pip install --upgrade joblib==1.1.0
```

A topic size (i.e., the number of documents containing the topic), and a topic numbers were created and display using the code below.

```
topic_sizes, topic_nums = model.get_topic_sizes()
print(topic_sizes)
print(topic_nums)
```

Because NLP algorithm produced lots of topic (i.e., 68) including the number of articles ranged from 29 to 549 among 8125 articles, only top 8 topics containing the largest numbers of articles were displayed using the code below.

```
topic_words, word_scores, topic_nums = model.get_topics(8)
for words, scores, num in zip(topic_words, word_scores, topic_nums):
  print(num)
  print(f"words:{words}")
```

The most similar topics are merged hierarchically to reduce the number of topics down to 4 from 68 after application of the code below.

```
topic_mapping = model.hierarchical_topic_reduction(num_topics=4)
```

The keywords belong to each broader topic (due to merging) is displayed through the following code.

`model.topic_words_reduced[3]`, wherein `[3]` indicates the index number of the broader topic.

The following code was applied to create and display the word cloud for each topic, where the font-size of the words aligned with the frequent of word in the documents belongs to each topic.

`model.generate_topic_wordcloud(3)`, wherein `[3]` indicates index number of topic.

3. Result

Relying on Top2Vec algorithm, topic modeling analysis has been conducted over 8125 article abstracts. The analysis yielded 68 topics. Size of each topic is presented in Table 1 below.

Table 1.

Topic Number and Corresponding Topic Size

	Topic Number and the Number of Documents									
Topic No	1	2	3	4	5	6	7	8	9	10
Number of Documents	549	438	381	377	257	228	223	204	192	188
Topic No	11	12	13	14	15	16	17	18	19	20
Number of Documents	182	180	178	172	165	164	159	146	146	145
Topic No	21	22	23	24	25	26	27	28	29	30
Number of Documents	144	141	132	132	131	131	129	114	110	110
Topic No	31	32	33	34	35	36	37	38	39	40
Number of Documents	109	104	96	91	86	84	78	77	71	68
Topic No	41	42	43	44	45	46	47	48	49	50
Number of Documents	67	66	65	63	61	59	58	58	56	56
Topic No	51	52	53	54	55	56	57	58	59	60
Number of Documents	50	50	48	46	43	43	42	41	40	37
Topic No	61	62	63	64	65	66	67	68		
Number of Documents	37	35	34	33	33	33	30	29		

This table shows the topic numbers and the number of articles matching the corresponding topic. For instance, topic number 1 possesses 549 documents from the dataset. In other words, 549 of the articles published in 2021 and indexed in Web of Science Core Collection database under the category of Education Scientific Disciplines can be grouped together since their content are closely related to each other. In another example, the topic number 60 contains 37 articles with similar content from the same dataset. As it can be noticed that the topics in the table as well as the output is ordered based on the number of article that each topic possess.

The sample keywords of the eight topics (the number of documents > 200) with the highest number of articles and the names given by the field experts are provided in Table 2 below whereas the word clouds of top eight topics created based upon word frequency is presented in Figure 1.

Table 2.

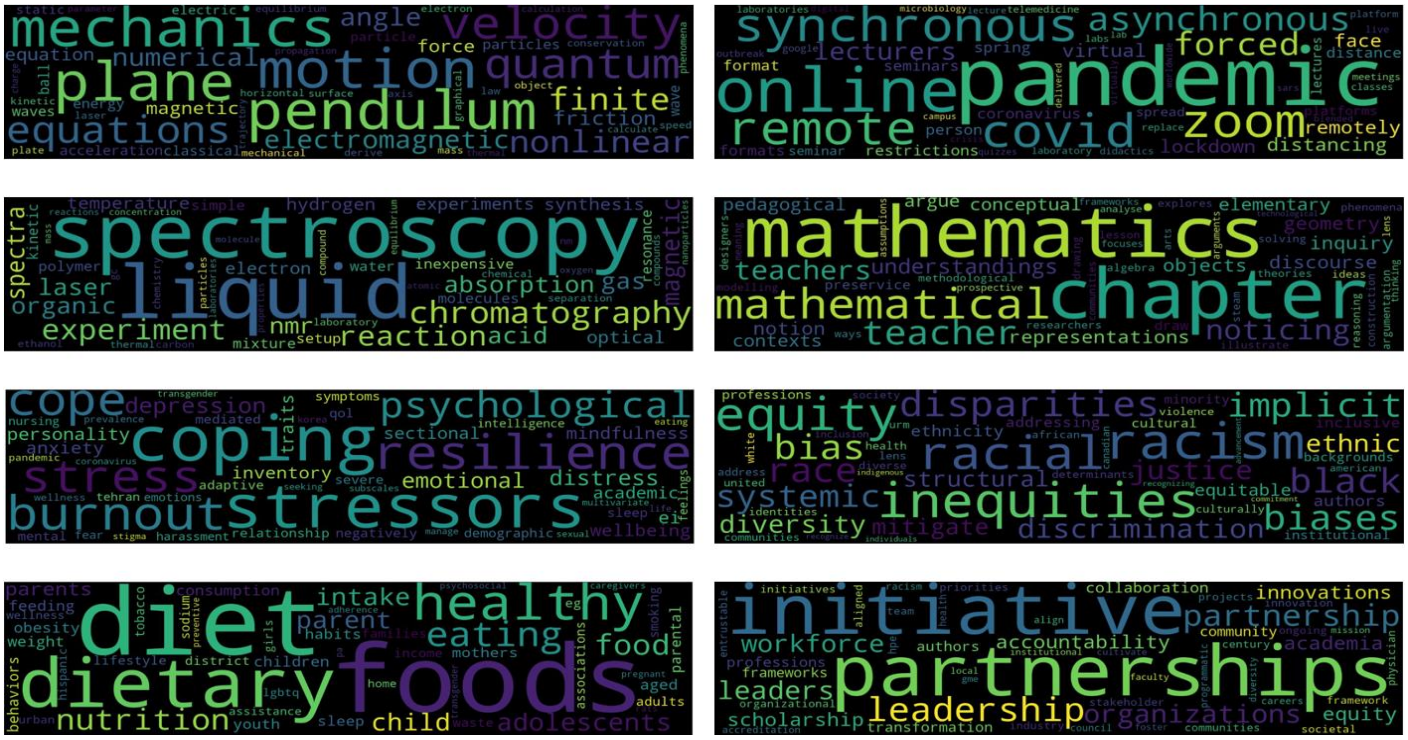
The Most Frequent Keywords, Sizes, And Given Names of The Largest Eight Topics

Topic No	Topic Size	Most frequent keywords within the topics	Topic name
1	549	'motion' 'pendulum' 'plane' 'mechanics' 'velocity' 'quantum' 'equations' 'nonlinear' 'finite' 'electromagnetic' 'numerical' 'angle' 'friction' 'magnetic' 'force' 'equation' 'particles' 'classical' 'ball' 'wave' 'acceleration' 'energy' 'waves' 'particle' 'static' 'electric' 'equilibrium' 'graphical' 'kinetic' 'speed' 'horizontal' 'object' 'mechanical' 'mass' 'surface' 'laser' 'axis' 'derive' 'conservation' 'phenomena' 'electron' 'law' 'plate' 'propagation' 'parameter' 'calculation' 'trajectory' 'charge' 'thermal' 'calculate'	Physics Education
2	438	'pandemic' 'online' 'synchronous' 'zoom' 'remote' 'covid' 'asynchronous' 'forced' 'lecturers' 'remotely' 'distancing' 'lockdown' 'virtual' 'face' 'restrictions' 'format' 'coronavirus' 'spring' 'seminars' 'formats' 'distance' 'person' 'platforms' 'seminar' 'lectures' 'spread' 'laboratory' 'telemedicine' 'laboratories' 'live' 'didactics' 'platform' 'outbreak' 'replace' 'meetings' 'labs' 'classes' 'google' 'microbiology' 'campus' 'virtually' 'quizzes' 'crisis' 'digital' 'lab' 'sars' 'delivered' 'worldwide' 'blended' 'lecture'	Online Learning and Covid
3	381	'spectroscopy' 'liquid' 'chromatography' 'experiment' 'reaction' 'absorption' 'laser' 'organic' 'nmr' 'spectra' 'magnetic' 'gas' 'acid' 'optical' 'hydrogen' 'temperature' 'synthesis' 'electron' 'experiments' 'inexpensive' 'simple' 'polymer' 'mixture' 'water' 'molecules' 'setup' 'kinetic' 'resonance' 'separation' 'laboratory' 'thermal' 'chemistry' 'compound' 'carbon' 'ethanol' 'gc' 'particles' 'compounds' 'chemical' 'nm' 'concentration' 'nanoparticles' 'mass' 'reactions' 'atomic' 'molecule' 'laboratories' 'properties' 'oxygen' 'equilibrium'	Chemistry Education
4	377	'mathematics' 'chapter' 'mathematical' 'teacher' 'noticing' 'teachers' 'understandings' 'representations' 'discourse' 'geometry' 'notion' 'inquiry' 'objects' 'pedagogical' 'elementary' 'conceptual' 'contexts' 'argue' 'preservice' 'draw' 'methodological'	Math Education and Reasoning

		'researchers' 'argumentation' 'modelling' 'steam' 'explores' 'phenomena' 'reasoning' 'ideas' 'ways' 'algebra' 'solving' 'illustrate' 'designers' 'lesson' 'construction' 'theories' 'drawing' 'prospective' 'arts' 'frameworks' 'focuses' 'assumptions' 'thinking' 'communities' 'lens' 'analyse' 'meaning' 'arguments' 'technological'	
5	257	'coping' 'stressors' 'resilience' 'burnout' 'stress' 'cope' 'psychological' 'depression' 'distress' 'emotional' 'personality' 'ei' 'mindfulness' 'traits' 'anxiety' 'inventory' 'wellbeing' 'academic' 'sectional' 'demographic' 'sleep' 'negatively' 'mediated' 'relationship' 'mental' 'severe' 'symptoms' 'qol' 'adaptive' 'intelligence' 'nursing' 'fear' 'feelings' 'tehran' 'harassment' 'emotions' 'pandemic' 'wellness' 'prevalence' 'Korea' 'tra*sgender' 'stigma' 'subscales' 'manage' 'eating' 'life' 'coronavirus' 'se*ual' 'seeking' 'multivariate'	Psychology and Emotions
6	228	'racism' 'inequities' 'equity' 'racial' 'disparities' 'implicit' 'biases' 'race' 'bias' 'black' 'systemic' 'discrimination' 'justice' 'ethnic' 'diversity' 'equitable' 'inclusive' 'addressing' 'address' 'authors' 'structural' 'mitigate' 'cultural' 'diverse' 'ethnicity' 'institutional' 'violence' 'white' 'policies' 'communities' 'health' 'united' 'medicine' 'populations' 'african' 'backgrounds' 'culturally' 'tra*sgender' 'minority' 'grounded' 'urm' 'recognize' 'mission' 'lg*tq' 'constructivist' 'identities' 'society' 'indigenous' 'persons' 'physicians'	Educational Diversity
7	223	'nutrition' 'healthy' 'foods' 'dietary' 'eating' 'diet' 'food' 'intake' 'parent' 'child' 'obesity' 'adolescents' 'aged' 'consumption' 'feeding' 'parents' 'mothers' 'weight' 'children' 'parental' 'youth' 'lifestyle' 'sleep' 'body' 'waste' 'behaviors' 'assistance' 'habits' 'income' 'tobacco' 'eg' 'adults' 'families' 'psychosocial' 'associations' 'serving' 'pregnant' 'sodium' 'district' 'caregivers' 'clinics' 'adherence' 'lg*tq' 'girls' 'rats' 'semistructured' 'home' 'intentions' 'survivors' 'centers'	Health and Life
8	204	'scholarship' 'mentoring' 'mentorship' 'mentors' 'scholarly' 'mentor' 'funding' 'publications' 'productivity' 'scholars' 'funded' 'publication' 'career' 'projects' 'journals' 'manuscript' 'academia' 'investigators' 'advancement' 'institutional' 'phd' 'networking' 'fellows' 'initiative' 'leadership' 'faculty' 'careers' 'authorship' 'journal' 'research' 'hpe' 'partnership' 'fellowship' 'leaders' 'annual' 'dissemination' 'reviewed' 'pursuing' 'indigenous' 'mission' 'institute' 'authors' 'pds' 'project' 'partnerships' 'urm' 'programs' 'pursue' 'residency' 'institutions'	Leadership and Mentoring

Figure 1.

Word Clouds of Top Eight Topics Created Based Upon Word Frequency



Top2Vec algorithm itself determines the number of the topics to be clustered. Since it is an unsupervised algorithm and trains the data set each time the Top2Vec model called, the number of topics would be slightly different even if it works on the same dataset. Therefore, the results of the topic modeling/ clustering (i.e., document size) would be slightly different each time. On the contrary, the results would be quite different when we force algorithm to reduce the number of topics. For instance, the document size in the first topic (index number is 0) in this study yielded 549 documents. The number of documents will be differed after topic reduction is applied because this procedure requires model to merge related topics together. In this study, I forced the model to reduce the number of topics from 68 to 4. Then, the model yielded the topic clusters with the following keywords:

Reduced Topic 1.

array(['plane', 'motion', 'liquid', 'laser', 'magnetic', 'pendulum', 'experiment', 'simple', 'particles', 'electron', 'nonlinear', 'finite', 'kinetic', 'chromatography', 'absorption', 'electric', 'optical', 'hydrogen', 'spectroscopy', 'equilibrium', 'temperature', 'electromagnetic', 'velocity', 'thermal', 'nanoparticles', 'particle', 'quantum', 'surface', 'angle', 'ball', 'friction', 'energy', 'mass', 'acid', 'mechanics', 'molecules', 'equations', 'resonance', 'ethanol',

'mechanical', 'reaction', 'compound', 'wave', 'experiments', 'acceleration', 'molecule', 'setup', 'antioxidant', 'graphical', 'gas'], dtype='<U15')

Reduced Topic 2.

array(['chapter', 'mathematics', 'chemistry', 'noticing', 'engineering', 'science', 'ideas', 'biology', 'inquiry', 'preservice', 'steam', 'arts', 'phenomena', 'mathematical', 'scientific', 'discourse', 'methodological', 'representations', 'solving', 'analyzes', 'designers', 'argumentation', 'understandings', 'classrooms', 'robotics', 'teacher', 'discusses', 'lesson', 'article', 'experimentation', 'computing', 'stem', 'teachers', 'computational', 'thinking', 'industrial', 'argue', 'hpe', 'engineers', 'algebra', 'decades', 'presents', 'geometry', 'arguments', 'explores', 'creative', 'researchers', 'scientists', 'introduces', 'societal'], dtype='<U15')

Reduced Topic 3.

array(['flipped', 'asynchronous', 'synchronous', 'anatomy', 'online', 'quizzes', 'osces', 'zoom', 'osce', 'stations', 'face', 'lecture', 'examiners', 'pandemic', 'format', 'summative', 'tutor', 'dissection', 'telemedicine', 'quiz', 'session', 'fc', 'station', 'lectures', 'lecturers', 'pocus', 'videos', 'video', 'tbl', 'examinations', 'anatomical', 'physiology', 'raters', 'blended', 'lockdown', 'remote', 'likert', 'examination', 'live', 'virtual', 'distancing', 'modality', 'marks', 'pathophysiology', 'radiology', 'tutors', 'ebm', 'neuroanatomy', 'exams', 'dental'], dtype='<U15')

Reduced Topic 4.

array(['tehran', 'tra*sgender', 'lg*tq', 'coronavirus', 'hpv', 'sectional', 'hospitals', 'council', 'globally', 'cme', 'united', 'crc', 'shortage', 'sought', 'hiv', 'burnout', 'vaccination', 'healthcare', 'pharmacist', 'pds', 'centers', 'ei', 'health', 'april', 'depression', 'dementia', 'wellness', 'concern', 'australia', 'adolescents', 'qol', 'entrustable', 'organizations', 'affiliated', 'palliative', 'care', 'purposive', 'workforce', 'harassment', 'iran', 'leave', 'accredited', 'background', 'fertility', 'se*ual', 'stigma', 'february', 'underserved', 'spiritual', 'obesity'], dtype='<U15')

4. Discussion and Conclusion

Top2Vec algorithm models the topics derived from huge volume of dataset. It is easy to apply since it does not require user to preprocess the text in dataset. It executes preprocessing automatically once the model employed. It reveals the optimum number of topics that the dataset content clustered into. This study revealed that the optimum topic size as 68. Among those topics, eight of then contained more than 200 publications indexed in Web of Science in 2021. Since the data analyzed in this study indexed in the Education Scientific Discipline

category of Web of Science citation index database, it is not surprising that the most studies fall into Physic Education, Math Education, and Chemistry Education categories. Among these, Online Learning and Covid-19 topic category contains the second largest number of document. Because Covid-19 pandemic obligated students, teachers/ instructors, managers, and educational policy makers to switch to online education; therefore compulsory online education caused by the pandemic became dominantly studied subject area. As the time this article is written, the fallowing query hit 108.000 results in google search engine: *Covid AND "online education" after:2020-12-31 before:2022-01-01.* This may be considered as evidence for why Online Education and Covid-19 is the second largest topic even in the Education Scientific Discipline category in WoS.

It is possible to preset the number of topics ordered by topic size to be displayed and reduce the numbers of topic into predetermined number by merging related topics. As it is demonstrated in the findings, when the topic size (the number of topics) was reduced down to 4 from 68, the coverage of each topic enhanced. If we take a closer look to reduced topic number 1, we can see some keywords such as 'liquid', 'particles', 'electron', 'hydrogen', 'temperature', 'molecules', 'ethanol', 'compound', 'molecule', 'antioxidant', and 'gas' did not belong to topic number 1 (i.e., Physics Education) before reduction. These keywords were in topic 3 (i.e., Chemistry Education) before topic reduction. Therefore, we may name the reduced topic 1 as 'Physics & Chemistry'. Similarly reduced topic 2 contains keywords some comes from some of the 68 initial topics. We may consider the name reduced topic 2 as 'STEM', which is stands for Science, Technology, Engineering, and Math. Naming the reduced topics 3 and reduced topic 4 is not as easy as naming the first two reduced topics since they contain keywords from diverse disciplines. Thus, deciding on the number of topics is crucial task that researchers should deal with. A researcher may try different topic sizes to gain optimum topic modeling performance.

5. Implications, Limitations, and Suggestions

Topic modeling may be useful tool for researchers, readers, authors, editors and so on. Topic models enhance their (i.e., researcher) ability to interpret data (i.e., research finding) by clustering them into different topics and classifying upcoming data into the most appropriate cluster (i.e., topic). Once we have research topics and subtopics, we can easily figure out what category should the upcoming research paper fall into. It would be beneficial for authors to locate the most related studies for literature review as well as to find appropriate journal to submit their manuscript. Likewise, editors may find topic models useful for classifying newly arrived manuscript into appropriate study topic which allow them to locate the authors conducted similar studies. Thus, the editors invite the authors of similar research publications as reviewer for newly arrived manuscript.

Limitation of this study may be lack of machine learning classification model to justify the usefulness of topics modeling for classification of newly arrived manuscript into appropriate topic category. Therefore, after topic modeling being conducted, the classification model best fits the data should be chosen among alternatives created and optimized applying diverse machine/ deep learning classification algorithms such as Random Forest, Naïve Bayes, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree etc.

Ethical Declaration

I declare that all scientific ethical rules were followed during the study.

Conflict Interest and Author Contributions

There is no conflict of interest. All stages of the study were organized and conducted by the Author(s).

6. References

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics.* Retrieved from https://arxiv.org/abs/2008.09470

Anuradha, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of student's performance. *Indian Journal of Science and Technology, 8*(15), 1-12.

Bohr, J.; Dunlap, R.E. (2018). Key topics in environmental sociology, 1990–2014: Results from a computational text analysis. *Environmental Sociology, 4*, 181–195.

Chang, I. C., Yu, T. K., Chang, Y. J., & Yu, T. Y. (2021). Applying Text Mining, Clustering Analysis, and Latent Dirichlet Allocation Techniques for Topic Classification of Environmental Education Journals. *Sustainability, 13*(19), 10856.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April). Topic modeling for evaluating students' reflective writing: a case study of pre-service teachers' journals. In Proceedings of *The Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5).

Egger, R., and Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers Sociology. 7*, 886498. doi: 10.3389/fsoc.2022.886498

Eykens, J., Guns, R., & Vanderstraeten, R. (2022). Subject specialties as interdisciplinary trading grounds: The case of the social sciences and humanities. *Scientometrics*, 1-21.

Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. (2020). Development of video-based emotion recognition using deep learning with Google Colab. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 18*(5), 2463-2471.

Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science 2015*, 349, 261–266.

Hung, J. L. (2012). Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. *British Journal of Educational Technology*, *43*(1), 5-16.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169-15211.

Karas, B., Qu, S., Xu, Y., & Zhu, Q. (2022). Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis. *Frontiers in Artificial Intelligence, 5*.

Linguamatics (2022). *What is Text Mining, Text Analytics and Natural Language Processing?* Retrieved (18.10.2022) from https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR), 9*, 381-386.

Mythili, M. S., & Shanavas, A. M. (2014). An Analysis of students' performance using classification algorithms. *IOSR Journal of Computer Engineering*, *16*(1), 63-69.

Sevli, O. (2019). Göğüs kanseri teşhisinde farklı makine öğrenmesi tekniklerinin performans karşılaştırması. *Avrupa Bilim ve Teknoloji Dergisi*, (16), 176-185.

Sevli, O., & Kemaloğlu, N. (2021). Olağandışı Olaylar Hakkındaki Tweet'lerin Gerçek ve Gerçek Dışı Olarak Google BERT Modeli ile Sınıflandırılması. *Veri Bilimi*, *4*(1), 31-37.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2007). *Data Mining In Excel: Lecture Notes and Cases*. Retrieved from https://www.researchgate.net/file.PostFileLoader.html?id=56fe049fcbd5c24f98534a44&assetKey=AS%3A345931493986304%401459487903102 (18.10.2022)

Zawacki-Richter, O., & Latchem, C. (2018). Exploring four decades of research in Computers & Education. *Computers & Education*, *122*, 136-152.

Zawacki-Richter, O.; Naidu, S. (2016). Mapping research trends from 35 years of publications in Distance Education. *Distance Education*, *37*, 245–269.