Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 29, S. 432-437, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 29, pp. 432-437, December 2021

Research Article

Research Trends Analysis in Educational Journal Publications on Covid19 Using Descriptive and Text Mining Methods: Preliminary Analysis

Cansu Çiğdem Ekin[*], Mustafa Çakıcı[1], Egemen Şener[1], Sıla Türker[1], Sinem Altanlar[1]

[*] Atılım Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye (ORCID: 0000-0003-4838-9708)
[1] Atılım Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2021 – 21-23 October 2021)

(**DOI:** 10.31590/ejosat.1036109)

ATIF/REFERENCE: Ekin, C. Ç., Çakıcı, M., Şener, E., Türker, S. & Altanlar, S. (2021). Research Trends Analysis in Educational Journal Publications on Covid19 Using Descriptive and Text Mining Methods: Preliminary Analysis. *European Journal of Science and Technology*, (29), 432-437.

Abstract

The study aims to reveal the studies' profile on covid19 in journals in the field of education. For this purpose, probabilistic topic modeling technique and decriptive analysis has been used to together to analyze 3039 journal articles that are indexed by the SCOPUS database between January 2020 and May 2021. Within the scope of decriptive analysis, the most cited journals, the most publishing journals, and the most publishing countries were analyzed. In probabilistic topic modeling stage, Latent Dirichlet allocation (LDA) algorithm which is a text mining method was applied to the abstracts of those extracted documents to identify topics in publications containing keywords such as covid, corona, pandemic in their titles. The results of text mining revealed 10 major topics mapping the the studies' profile on covid19 in journals in the field of education. In this study, preliminary analysis results were given.

Keywords: COVID-19, Latent Dirichlet Algorithm, Online Education, e-Learning, Digital Education, Topic Modeling

Betimleyici ve Metin Madenciliği Yöntemleri Kullanılarak Covid19 Konulu Eğitim Dergisi Yayınlarında Araştırma Eğilimleri Analizi: Ön Analiz

Öz

Çalışma, eğitim alanındaki dergilerde yer alan Covid19 ile ilgili yapılan çalışmaların profilini ortaya çıkarmayı amaçlamıştır. Bu amaçla, Ocak 2020 ile Mayıs 2021 tarihleri arasında SCOPUS veri tabanı tarafından indekslenen 3039 dergi makalesini analiz etmek için olasılıksal konu modelleme ve betimsel analiz ile birlikte kullanılmıştır. Betimsel analiz kapsamında en çok atıf alan dergiler, en çok yayın yapan dergiler ve en çok yayın yapan ülkeler analiz edilmiştir. Olasılıksal konu modelleme aşamasında ise; başlıklarında covid, corona, pandemi gibi anahtar kelimeler içeren yayınlardaki çalışma konularını belirlemek için ilgili belgelerin özetlerine Latent Dirichlet Tahsisi (LDA) algoritması uygulanmaktadır. Metin madenciliği sonuçları, eğitim alanındaki dergilerde covid19 ile ilgili çalışmaların profilini haritalayan 10 ana konuyu ortaya koymuştur. Bu çalışmada ön analiz sonucları verilmiştir.

Anahtar Kelimeler: COVID-19, Latent Dirichlet Algorithm, Online Eğitim, e-Learning, Digital Eğitim, Konu Modelleme

[*] Cansu Çiğdem EKİN, cansu.aydin@atilim.edu.tr

1. Introduction

The power of the Internet, information and communication technologies, the rapid accessibility in data networks, and have caused great changes in social life, health sector, education, economy and so on. E-learning is a learning method that emerged with the use of modern communication and internet technologies. It is possible to access information from anywhere and anytime by using e-learning tools. In this way, it continues and improves the existing traditional education programs.

The covid-19 outbreak education, which is currently on the agenda, had effects in some ways. The COVID-19 outbreak has been spreading all over the world since the day it emerged. This epidemic process does not only affect health systems and their economic dimensions, but also education systems. The pandemic process has led to the formation of concepts such as distance education, online education, hybrid education on education, etc. When the literature were analysed for researches that use the text mining method on covid-19, it was seen that the concept of text analysis, which is used synonymously with text mining, was used to draw a quantitative result by examining patterns and trends in different data sets. Text analysis, such as sentiment analysis, content analysis, keyword analysis, topic analysis can be used to find out meaningful information from the data (Yang & Zhang, 2018). These analysis methods have also been used in recent studies. Isoaho ve his colleques (2019) used topic modeling and text analysis to review qualitative policy research. Kim and Lee (2019) performed a network text analysis of medical tourism in newspapers using text mining. Bi and his colleques (2018) made a systematic mapping using text mining in software architecture. In the scope of this study, articles on covid-19 were examined. Considering that we only have 2-year data on Covid-19, most of these studies have been conducted on the evaluation of Covid-19 in terms of health (Tworowski et al., 2021; Glowacki et al, 2020), its impact on the population (Glowacki et al, 2020; Tworowski et al., 2021), its impact on the industry (Atay et al, 2021; Yang& Han, 2021), and how it affects people's psychology (Lyu&Luli, 2021; Koh&Liew, 2020).

There is no text mining work done in the education field of Covid-19. Our work is the first in this area. The reason why we use the articles about covid-19 in the field of education in this analysis is that in most of the world, education is transferred online and e-learning patterns are integrated into our lives.

As a result, using the text mining technique, articles on Covid-19 are discussed in areas such as people's health, people's psychology, effects on the economy and industry. Our project applies text mining technique on articles written in the education field of covid-19. The fact that no previous work has been done in this area is a work of our project to fill the gap in this area.

2. Background

2.1 LDA Model

The algorithms suggested for subject modeling are statistical methods and aim to reach a conclusion by analyzing the words that make up the document. Latent Dirichlet Allocation (LDA), which has gained great importance in machine learning and text mining applications and is one of the most basic and popular subject modeling methods, is a generative graphical model used to model discrete data such as documents and find reveal the topic of the document. Each document is made up of a random mix of topics, and each word that makes up the document is chosen from a topic (See Figure 1). Topics also show a probability distribution from words in a fixed dictionary. The process first starts with sampling the words in the dictionary under topics. The next step illustrates the possibility for each topic to be included in the document. The probability of locating words under topics and subjects in the document is obtained by the Dirichlet distribution. In topic modeling methods, the number of subjects is determined initially by assuming that they are known.

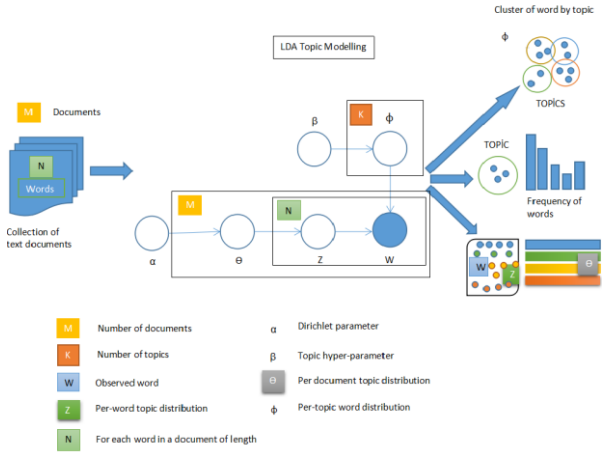


Figure 1. LDA Topic Modelling Process

In topic modeling methods, the number of subjects is determined initially by assuming that they are known. Coherence measure value allows us to score a single subject by measuring the degree of semantic similarity between words with a high score on the subject.In this study, C_v measure was used as a consistency criterion for performance comparison.*C v* measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity [21].The number of topic with the highest consistency value was determined as the number (K) of which we will train our system. In this study, the number of topics was accepted as 10.

Also, subject modeling methods are not sensitive to Dirichlet parameters. In this study, Dirichlet parameters were assigned symmetrically and alpha value was determined as 0.1, and beta value was determined as 0.01. The Alpha parameter is the Dirichlet parameter that represents the distribution of the topics in the documents.A higher alpha value indicates that documents have more topics, resulting in a more specific topic distribution per document.The beta parameter is the parameter that represents the distribution of words in topics.A high beta value indicates that it results in a more specific word distribution per topic.

3. Methodology

3.1 Data Collection

The data were obtained using the Scopus database. The reasons why Scopus database is preferred when obtaining data; its connections it has a lot of features, such as presenting author profiles that include the number of publications and bibliographic data, references and details about the number of citations each

published document has received(See Figure 2). While obtaining data from the Scopus database, predetermined keywords were used and the articles containing these words were listed.

Medical resources are not included among these listed articles and are limited to educational resources. The listed articles are generally composed of articles published in English. Due to the examination of the education field of the Covid-19 pandemic, the research has been restricted with articles published only in the last two years.

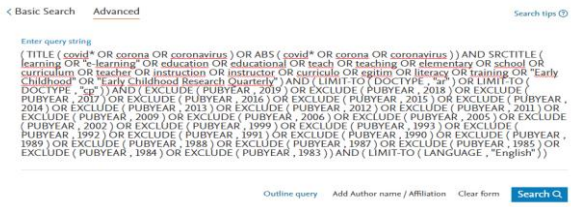


Figure 2. All words and excluded features

3.2 Working Environment

Coherence scores of topic models were calculated using the Gensim (ver. 4.0.1) library on Python version 3.9 programming language, and they were run in an environment with Windows 10, Intel i7 processor and 16 GB ram. Topic results and analyzes were made via Google Colab.

3.1 Preprocessing

The input data to be analyzed in data mining must have a certain format and must be cleared of corrupt or unnecessary data. The biggest problem with text mining is that the data set it will process is not structured. The pre-processing stage in the field of text mining, which is generally working on documents written using natural language, performs the process of converting the data to the appropriate format in addition to data cleaning. In this study, the Natural Language Toolkit library created for python is used [18]. NLTK includes modules related to text processing and analysis algorithms that enable raw text data to be processed.

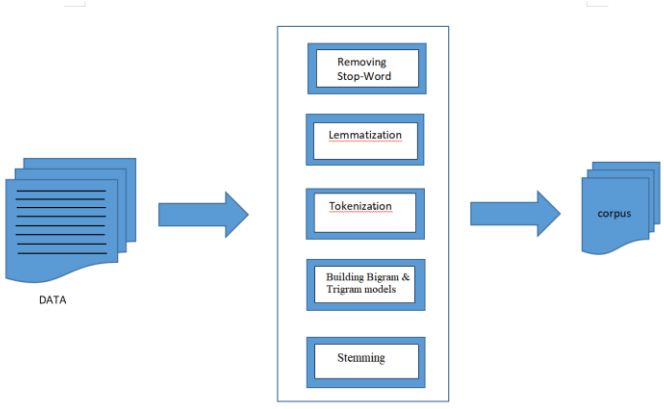


Figure 3. Steps of Data Preprocessing

Text preprocessing steps generally consist of five steps as stated below (see Figure 3) ;

1. Split text into sentences (Tokenization)
2. Filter out unnecessary words (stopwords filtering)
3. Word body (lemmatization): Finding the main forms of the words and converting the word into body form
4. Building Bigram & Trigram models: Distributed representations of words in a vector space help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words [19]. Bigrams are two words that are frequently occurring together in the document. Trigrams are three words that are frequently occurring together in the document.
5. Stemming: It is a root-finding algorithm that allows finding the root of a word by normalizing it linguistically. The goal of the root-finding algorithm is; to reduce the variety of words studied by rescuing the words with inflectional or constructional suffixes from the suffixes.

The Gensim library was used to perform topic modeling on the data that was cleared after preprocessing [20]. Gensim is designed to process large text collections using data flow and incremental online algorithms; This feature distinguishes it from many other machine learning software packages that only target in-memory processing, for this reason, Gensim has been used and cited in many various disciplines, commercial and academic applications, from medicine to insurance, claims analysis and patent research. In text mining, the regular and structural coexistence of the cleaned data is called corpus. The LDA algorithm in the Gensim library was used to examine the topics discussed in these texts, the recurring themes and the extent to which each document deals with these topics through the corpus we created.

4. Results of the Study

4.1 Descriptive Analysis

We extracted 3.039 documents that were published in academic journals and conference papers, listed in the Scopus database. Of those documents, 2.825 were journal articles. Before applying LDA to these 3.039 articles, descriptive statistics are obtained based on the distribution of related documents and sources.

The 3.167 documents were published in 679 sources. Table 1 lists the annual frequencies of documents for the top 10 sources. As Table 1 shows, Journal Of Chemical Education Including Subseries Teaching Chemistry in the Time of COVID-19 and Experimenting with At-Home General Chemistry Laboratories during the COVID-19 Pandemic publishes far more documents in this field than do any of the other top sources.This source focuses on Attempts, successes, and failures of distance learning in the time of covid-19 and challenges of laboratory teaching with the challenges of Covid-19.

Table 1. Top 10 sources of articles on education field of the Covid-19 pandemic.

#	Source	# of documents
1	Journal Of Chemical Education	184
2	Education Sciences	73
3	Journal Of Dental Education	68
4	Gms Journal For Medical Education	63
5	Frontiers In Education	55
6	Journal Of Microbiology And Biology Education	54
7	BMC Medical Education	53
8	Journal Of Surgical Education	50
9	Education And Information Technologies	49
10	Biochemistry And Molecular Biology Education	40

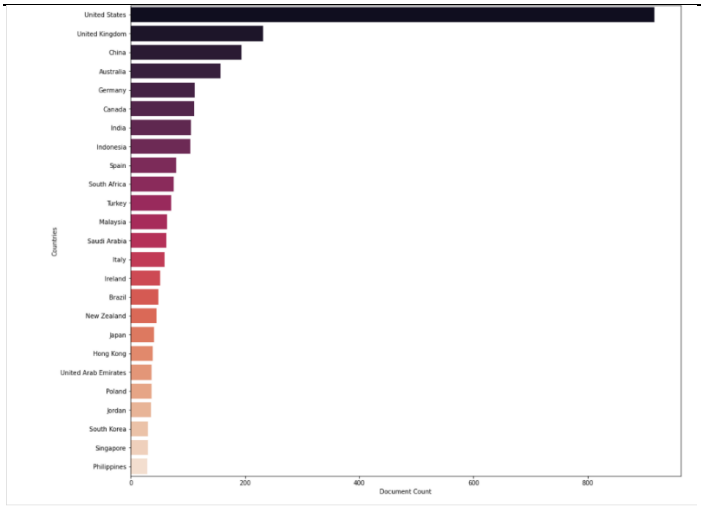


Figure 4. Number of Documents by Countries

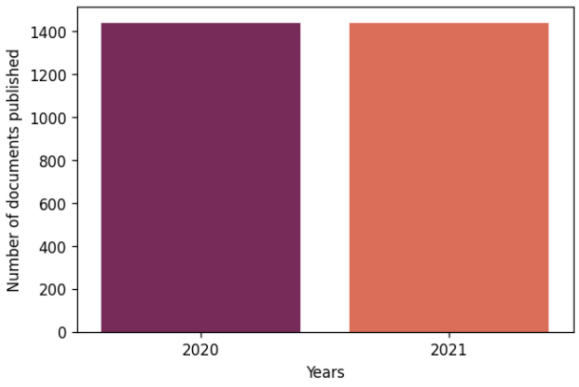


Figure 5. Number of Documents by Year

In figure 4, the first 25 countries where studies examining the impact of covid-19 on the field of education have been published are shown with a column chart. The United States is at the top of the list as the country that broadcasts the most. It is followed by the United Kingdom, China, and Australia.

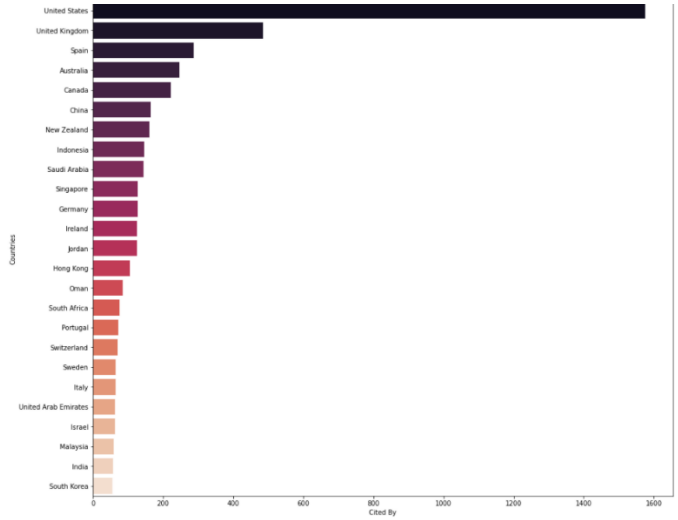


Figure 6. Number of Citations by Countries

In figure 6, the first 25 countries according to the number of citations of the countries where studies examining the impact of covid-19 on the field of education are published are shown with a column chart. The United States tops the list as the country with the most citations. It is followed by the UK, Spain, Australia, and Canada.

The spread of covid-19's studies in the field of education by years is shown in Figure 5. Since covid-19 entered our lives 2 years ago, we have data showing only 2 years. In general, documents were published at very close rates every 2 years.

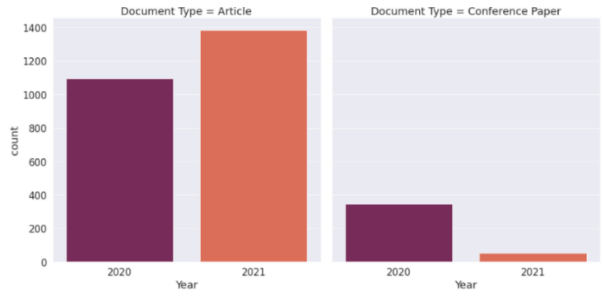


Figure 7. Number of documents by type and year

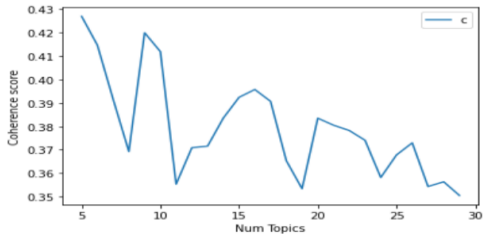
Most of the published documents consist of articles. As can be seen from the Figure 7, more articles were published in 2021 than in 2020. Conference papers are less than articles. Conference papers published in 2020 are more than in 2021. However, it is important to remember that 2021 continues now. These data may change at the end of the year.

4.2 Topic Modeling Analysis

4.2.1 Topic modeling analysis with Gensim

For more specific research trends, text mining is used on the source title, abstracts and keywords that describe research on education field of the Covid-19 pandemic. A dictionary is imported to extract the appropriate lemma through morphological

analysis. To select principal terms in sequence, stop words and words with fewer than two letters are removed, and terms that appear frequently in many documents are weighted. Figure 8. shows the results of using the coherence score value as a measure to select the optimal topics. From the highest coherence score value, we determined that the optimal number of topics for the model is 10. Dirichlet parameters were assigned symmetrically and alpha value was determined as 0.1, and beta value was determined as 0.61.




```
Num Topics = 5  is having Coherence Value of 0.4269198966698067
Num Topics = 6  is having Coherence Value of 0.41461966124996685
Num Topics = 7  is having Coherence Value of 0.39156236021816765
Num Topics = 8  is having Coherence Value of 0.369227134572336
Num Topics = 9  is having Coherence Value of 0.4199334576841103
Num Topics = 10  is having Coherence Value of 0.41188050384806907
Num Topics = 11  is having Coherence Value of 0.3553154210997256
Num Topics = 12  is having Coherence Value of 0.3708594607926469
Num Topics = 13  is having Coherence Value of 0.37154481195212213
Num Topics = 14  is having Coherence Value of 0.38353476759836924
Num Topics = 15  is having Coherence Value of 0.39238575602722153
```

Figure 8. Changes in coherence score by number of topics.

Table 2. Determining the Contents of the Study with the first ten words

Topic #	Content of the Studies	Topic Keywords
0	studies related with interview session results for virtual simulation in medical curriculum for clinical skills	Virtual, simulation, student, medical, interview, session, clinical, skill, curriculum, format
1	studies related with experiences of students/patients/residents in medical training program during pandemic	Medical, training, clinical, program, patient, resident, care, experience, student, conclusion
2	studies related with student online learning in university	Online, student, learn, teacher, study, education, teaching, university, learning, teach
3	studies related with educational for supporting educational program challenge during pandemic	Education, school, practice, pandemic, support, student, program, challenge, leadership, educational
4	studies related with school experiences	School, child, teacher, parent, pandemic, education, experience, study, support, learn
5	studies related with examination method in medical exam for student online assessment	Student, online, assessment, exam, medical, time class, method, examination, teaching
6	Studies related with deep learning	Model, image, base, video, learn, detection, algorithm, paper, propose, network
7	studies related with educational and social policy during pandemic on impact of community health crisis	Education, social, pandemic, policy, community, crisis, health, paper, impact, article
8	survey studies related with knowledge level and factors influences student mental health during pandemic	Student, study, health, level, school, pandemic, knowledge, survey, covid, factor
9	studies related with student online learning experiences, educational challenges teaching remote	Student, learn, learning, online, experience, teaching, remote, virtual, Challenge, education

In Table 2, we list the best coherence score in 10 topics names, dominant word weight of the topic, and the topic keyword for each topic. The topic name is defined by manual examinations based on our prior knowledge and extracted terms. The process of finding the number of topics and defining the topic name was done by the field expert in education who is an expert also in the field of text mining. In most cases, the first five keywords were combined in a meaningful manner to name each topic. As seen in Table 2, first topic with the highest weight includes studies related with examination method in medical exam for student online assessment. Second topic includes studies related with student online learning in university. Third topic is related with studies about student online learning experiences, educational challenges teaching remote. Following topics are related with studies on student online learning in university, experiences of students/patients/residents in medical training program during pandemic and studies on school experiences.

5. References

Atay, M., Eroğlu, Y., & Ulusam Seçkiner, S. (2021). Investigation of Breaking Points in the Airline Industry with Airline Optimization Studies Through Text Mining before the COVID-19 Pandemic. Transportation Research Record, 0361198120987238.

Bi, T., Liang, P., Tang, A., & Yang, C. (2018). A systematic mapping study on text analysis techniques in software architecture. *Journal of Systems and Software*, *144*, 533-558.

Glowacki, E. M., Wilcox, G. B., & Glowacki, J. B. (2021). Identifying# addiction concerns on twitter during the COVID-19 pandemic: A text mining analysis. Substance abuse, 42(1), 39-46.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. Policy Studies Journal, 49(1), 300-324.

Kim, S., & Lee, W. S. (2019). Network text analysis of medical tourism in newspapers using text mining: The South Korea case. Tourism Management Perspectives, 31, 332-339.

Koh, J. X., & Liew, T. M. (2020). How loneliness is talked about in social media during COVID-19 pandemic: text mining of 4,492 Twitter feeds. Journal of psychiatric research.

Lyu, J. C., & Luli, G. K. (2021). Understanding the public discussion about the centers for disease control and prevention during the covid-19 pandemic using twitter data: Text mining analysis study. Journal of Medical Internet Research, 23(2)

Tworowski, D., Gorohovski, A., Mukherjee, S., Carmi, G., Levy, E., Detroja, R., ... & Frenkel-Morgenstern, M. (2021). COVID19 Drug Repository: text-mining the literature in search of putative COVID19 therapeutics. Nucleic acids research, 49(D1), D1113-D1121.

Yang, M. & Han, C. "Revealing industry challenge and business response to Covid-19: a text mining approach", 2021.

Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. International Journal of Computer and Information Engineering, 12(7), 525-529.