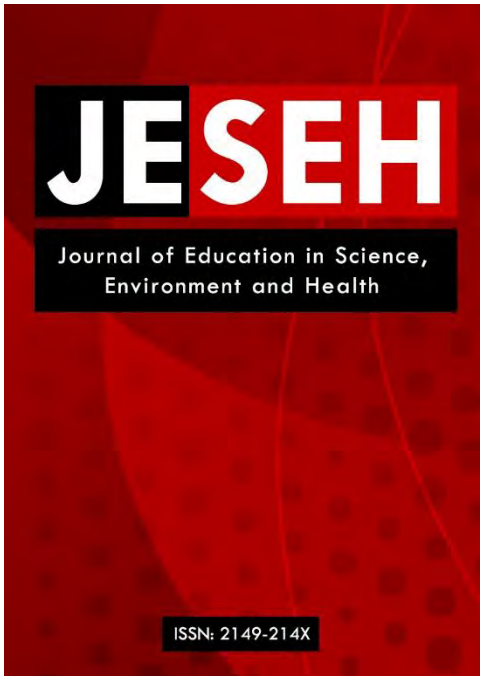

ISSN: 2149-214X

Journal of Education in Science, Environment and Health

www.jeseh.net

Using Artificial Intelligence to Analyze Nature of Science Themes: Einstein and Eddington Documentary Film Example

Yusuf Kartal, Munise Seckin-Kapucu
Eskisehir Osmangazi University

To cite this article:

Kartal, Y. & Seckin-Kapucu, M. (2024). Using artificial intelligence to analyze nature of science themes: Einstein and Eddington documentary film example. *Journal of Education in Science, Environment and Health (JESEH), 10*(4), 196-207. https://doi.org/10.55549/jeseh.710

JESEH
Journal of Education in Science, Environment and Health
Volume 10, Issue 4, 2024
https://doi.org/10.55549/jeseh.710

Using Artificial Intelligence to Analyze Nature of Science Themes: Einstein and Eddington Documentary Film Example

Yusuf Kartal, Munise Seckin-Kapucu

Article Info

Article History

Published:
01 October 2024

Received:
10 March 2024

Accepted:
29 July 2024

Keywords

Nature of science
Natural Language Processing
Latent Dirichlet Allocation
LDA
Topic Modelling

Abstract

Understanding the nature of science is an essential component of scientific literacy. In a technology and media-oriented environment, text-processing algorithms and various artificial learning approaches are crucial and continue to develop. Latent Dirichlet Allocation is a topic modeling algorithm that has been used frequently for many years to extract the main themes in many documents. This study examined the nature of science themes in the documentary film "Einstein and Eddington," which lasted 1 hour and 28 minutes, with the Latent Dirichlet Allocation topic extraction algorithm. First, the texts in the documentary film were fragmented into 30-second, 1-minute, and 3-minute periods to obtain three different datasets. Considering the literature on these datasets, experiments were carried out using the Latent Dirichlet Allocation algorithm to extract 5, 7, and 9 topics. The Latent Dirichlet Allocation algorithm developed with R programming language was used to analyze the data. In the analysis made by the computer, it is seen that science-related words such as science, scientist, theory, the universe, Eddington, think, speed, and answer stand out. While it was observed that it was difficult to distinguish the detected topics from each other on limited data, it was concluded that the dataset created with 30-second periods made more sensitive topic inferences. Despite the challenges posed by subjectivity regarding the nature of science, it is thought that computer-aided models can provide much more than information retrieval and advanced search. In this context, although it seems pretty difficult to extract the nature of science themes on limited data with the Latent Dirichlet Allocation algorithm, it is possible that the artificial learning models to be used in addition to the Latent Dirichlet Allocation algorithm can detect the nature of science themes.

Introduction

21st-century individuals live in a technology and media-oriented environment. In this environment, individuals use many applications or devices containing advanced technology. Commonly used and developing technologies include augmented reality, virtual reality, mixed reality, mobile applications, web 3.0/4.0, cloud technology, simulation, social networks, educational and digital games, digital storytelling, artificial intelligence, online learning environments, wearable technology, QR Code applications, three-dimensional printing, etc. (Adams Becker et al., 2016; Johnson et al., 2015). These technologies are developing under the leadership of science and reaching individuals.

Societies are increasingly confronted with science and technology, which is its application, day by day. This situation arises from the constantly developing and changing structure of science. Understanding science is necessary to see the results of the application of science and to recognize scientific thought. This is an intellectual necessity for the intellectuals of our age (Yıldırım, 2010). Knowing science and its characteristics is also of critical importance for scientific literacy. According to Shen (1975), scientific literacy has three dimensions: practical, civil, and cultural. Practical scientific literacy focuses on solving the daily problems of life, civic scientific literacy focuses on issues related to the protection of the natural environment, and cultural scientific literacy focuses on knowing the ideas that represent outstanding cultural achievements in science. According to Hodson (2008), in order to understand scientific literacy, it is necessary to understand the nature of science (NOS), understand the basic theoretical frameworks of biology, chemistry, and physics, understand the complex relationships between science, technology, society, and environment, to be aware of contemporary applications of science, to have the ability to use science in solving daily problems. Having a personal opinion on controversial issues from a scientific and technological perspective and a basic understanding of global environmental issues is vital. It is advocated that the NOS should be taught in order to increase scientific literacy (Hand et al., 1999).

It is possible to define the NOS as the way of accessing knowledge or the values and beliefs inherent in developing scientific knowledge (Abd-El-Khalick et al., 1998; Lederman, 1992). Understanding the NOS is seen as an essential component of science literacy (Lederman, 2007) and is emphasized in many reform reports (American Association for the Advancement of Science [AAAS], 1989, 1993; National Science Teachers Association [NSTA], 1982; National Research Council [NRC], 1996).

McComas et al. (1998) explore how science operates, the dynamics of scientists as a social unit, societal influence on and reactions to scientific pursuits, and different facets of social sciences like philosophy of science, history of science, and sociology of science. They view it as a fusion of cognitive sciences like psychology, which delve into interpretations. According to Lederman (1992), the NOS is seen as the values and beliefs inherent in scientific knowledge. Although there is no single, universally accepted definition of the NOS, there is now a significant academic consensus on what aspects of the NOS should be taught in schools (Lederman et al., 2002; Smith et al., 1997; Smith & Scharman, 1999). As a result of scientific studies, it has been determined that some dimensions are necessary for teaching the NOS in formal education (McComas, 1998). These dimensions have been stated as follows. Scientific knowledge is not specific, is based on experiments, is subjective, is partly a product of human imagination and creativity, and is affected by the social and cultural environment. In addition, observation and inference and scientific theories and laws are different types of information (Lederman, 2007; Lederman et al., 2002; McComas, 1998; Smith & Scharman, 1999).

The dimensions of the NOS are not adequately understood by students, scientists, teachers, and teacher candidates (Irez, 2006; Ryan & Aikenhead, 1992; Dagher & Boujaoude, 2005). Many people have misconceptions about the NOS. The sources of these common misunderstandings are textbooks, teachers, and the media (McComas, 2003; Irez, 2006). In the literature, there are studies on films and film genres in the field of education (Başkalyoncu, 2017; Dark, 2005; Efthimiou & Llewellyn, 2006; 2007; Navakanesh et al., 2019; Piliouras et al., 2011; Yıldırım et al., 2015; Yılmaz, 2018). However, some studies examine documentary films, which are a means of transferring science, and their content in terms of the concepts of the NOS (Başkalyoncu, 2017; Seckin-Kapucu et al., 2015; Seckin-Kapucu, 2016; Seckin-Kapucu, 2023). When studies on the NOS are examined, studies on the Latent Dirichlet Allocation (LDA) model attract attention but are limited (Allen & Murdock, 2020; Wang et al., 2023). In a study by Allen and Murdock (2020), applying LDA topic modeling to questions of interest to historians and philosophers of science was discussed and exemplified through Charles Darwin's work on modeling reading and writing behavior. In a study conducted by Wang et al. (2022), prospective teachers' understandings of the NOS were examined with the LDA model.

LDA is a widely used topic modeling algorithm applied to textual data, which forms the basis of many topic extraction models (Li & McCallum, 2006, pp. 577-584). Topic modeling is a common approach used for text processing. It is frequently used to reveal the main themes in documents (Yang & Zhang, 2018). LDA is a statistical method used to identify abstract topics embedded in a set of documents (Blei et al., 2003; Wang et al., 2018). On the other hand, LDA is an effective method that stands out in topic modeling (Zhao et al., 2020). LDA discovers hidden themes in documents. It uses a generative probability model and Dirichlet distributions (Kaya & Gülbandılar, 2022). According to the LDA algorithm, all words represent a topic to a certain extent. All documents contain these topics to a certain extent so that each document can be expressed as a mixture of more than one topic. In other words, in LDA analysis, documents are seen as a component of different topics (Calvo et al., 2018; Wang et al. 2023). LDA enables the detection of topics in the entire archive by calculating the probability of a topic being in a document (Kartal, 2017). There are various studies in the literature using the LDA model. When these studies are examined, Facebook, Twitter, Blog, etc. It is seen that social media contents such as these are analyzed (Güven et al., 2018; Lin & He, 2011; Pak & Paroubek, 2010; Roberts et al., 2012; Tur et al., 2013). Lin and He (2009) simultaneously extracted product features and emotional expressions from cinema reviews with the JointSentiment topic model, an LDA-based method they developed.

Many methods are used to determine which NOS theme documentary content belongs to. In this study, computer analysis of the NOS themes in the science content presented, primarily through the media, has been a matter of curiosity. In this study, the LDA algorithm, one of the leading topic modeling techniques, was preferred to detect the nature of the science theme of documentary content. In this context, the study on the film Einstein and Eddington, made by Seckin-Kapucu (2016), was requested to be analyzed with artificial learning and analysis methods regarding the elements of the NOS. In this context, the results obtained in the computer environment were examined using the LDA model. This study aimed to examine the NOS themes in the documentary film "Einstein and Eddington" according to the LDA Model and to extract the NOS themes in the documentary film. Within the scope of this general purpose, answers were sought to the following questions:

- What are the results of the analysis performed with LDA to extract topics on “Einstein and Eddington Documentary Film”?
- What are the results of the analysis performed with LDA at a dataset divided by 30-second, 1-minute, and 3-minute intervals?

Method

Model of the Research

In this study, LDA, a probability-based topic modeling method, was used. The model reveals the semantic structure of the text document by basic topic modeling with the words and weight values formed by the observed dataset (Güven et al., 2018). Topic modeling technology emerges as a new method for large-scale literature research (Yin & Yuan, 2022). As a natural language processing method, topic modeling can discover hidden, valuable, and relevant topics in large amounts of textual data (Jiang et al., 2016; Zhou et al., 2021). This model was preferred in this study because the dialogues in the Einstein and Eddington documentary film needed to be examined in terms of the NOS.

Data Source

The documentary film "Einstein and Eddington," which is 1 hour and 28 minutes long, was used as the data source in the study. Einstein and Eddington is a 2008 British drama produced by Company Pictures and the BBC in association with Home Box Office (HBO). The film was directed by Philip Martin. The film tells the story of Einstein's theory of general relativity, his relationship with Eddington, and introducing this theory to the World against the background of the Great War and Eddington's eclipse observations. The conversations in the Einstein and Eddington documentary film constitute the text data to which the method used in this study is applied. These text data were organized in 3 different ways to obtain three different datasets to work on (Table 1). The periods for the datasets are 30 seconds, 1 minute, and 3 minutes. The numbers of records are 174, 87, and 29 depending on the period.

Table 1. Information on the period and number of records of datasets obtained from the texts of the Einstein and Eddington documentary film

	Period	**Record Count**
Dataset-1	30 seconds	174
Dataset-2	1 minute	87
Dataset-3	3 minutes	29

In the referenced documentary (Seckin-Kapucu, 2016), five themes of the NOS were emphasized. For this reason, LDA was used to extract five topics in this study. However, to expand this and measure the general topic extraction abilities of the LDA algorithm, a study was conducted to identify seven topics based on the themes stated by the common opinion of experts in literature (Lederman et al., 2002; Smith et al., 1997; Smith & Scharman, 1999). In addition, considering that different topics could be included, a study was carried out to identify nine topics.

Process

In the study conducted on the documentary film Einstein and Eddington (Seckin-Kapucu, 2016), the themes of the NOS were directly examined by the researcher, and transcripts of the conversations in the film were obtained for these analyses. By examining these transcripts, the NOS themes were revealed. In order to measure the ability of LDA topic modeling to extract the NOS themes within the scope of this study, these transcripts obtained by Seckin-Kapucu (2016) were divided into three different formats, 30 seconds, 1 minute, and 3 minutes apart, and documents were created (Table 1). In this way, three datasets of different sizes containing the same data were obtained. The LDA topic modeling algorithm was run in 3 different iterations to extract 5, 7, and 9 topics from these datasets, referring to the NOS themes. In this way, nine different combinations were obtained on the same data. By examining the results obtained with the LDA algorithm, the theme of the NOS to which the keywords suggested by LDA might correspond was examined (Figure 1).

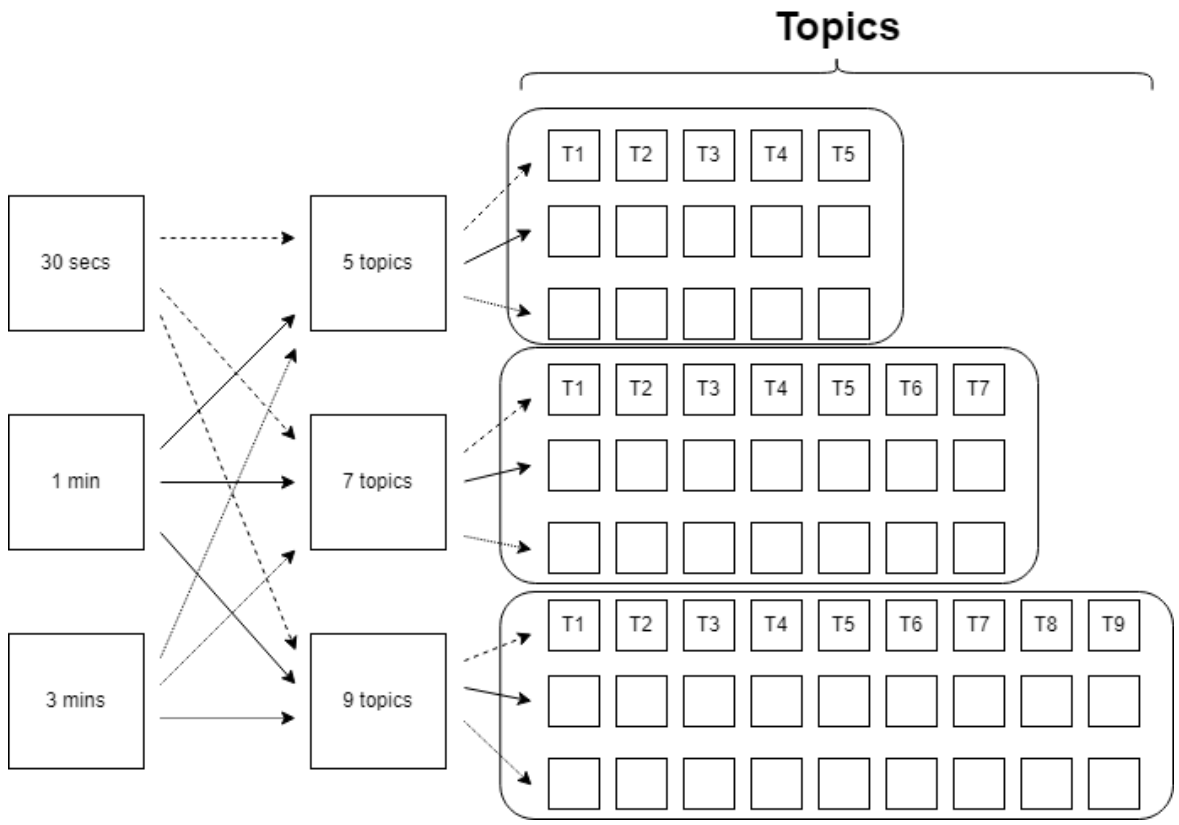

Figure 1. Process steps operated by LDA

Data Analysis

R programming language was used to analyze the data of this study. R programming language is a statistical software development library with functions that work as open-source scripts (Hornik, 2012). This language, developed in the Department of Statistics at Auckland University, New Zealand, has found widespread use over time (Ihaka & Gentleman, 1996). Unlike other languages, R provides access to data through references, thanks to unique data structures, rather than providing direct access to data (Özdemir et al., 2010). Since it is open source, it has function libraries developed by many users from different fields. Since it works as a script, it is compatible with all operating systems (Çelik, 2018). With the R programming language, nine different processes were run separately for the LDA algorithm to extract five, seven, and nine topics on each dataset.

The LDA function provided by the R programming language was used. The parameters of the LDA function are given in Table 2. *Gibbs Sampling* is chosen as the method for the LDA function. Gibbs sampling is one of the Markov Chain Monte Carlo (MCMC) methods used to sample from high-dimensional probability distributions. Gibbs sampling is beneficial when it is difficult to test all variables of complex probability distributions simultaneously. The method works by sampling each variable sequentially and from a conditional probability distribution.

Table 2. LDA Parameters

Parameter Name	**Parameter Value**
Method	Gibbs
control.estimate.beta	TRUE
control.seed	2138
control.nstart	1L
control.best	TRUE
control.var.iter.max	20
control.var.iter.tol	1e-6
control.em.iter.max	1000
control.em.iter.tol	1e-3
control.initialize	random

Results

It was determined that the film examined by Seckin-Kapucu touched on five themes related to the NOS. These are: Scientific knowledge is variable; Scientific knowledge involves logical, mathematical, and empirical inferences; Scientific knowledge is subjective; Imagination and creativity play an essential role in obtaining scientific knowledge; Scientific knowledge is affected by the social and cultural environment during its development and translation into practice. The NOS themes in the movie "Einstein and Eddington," which lasts 1 hour and 28 minutes, are explained by giving examples from the conversations in the movie (Table 3). When the words in each subject in the Datasets are examined, they have different frequency values. The resulting word sets are presented as word clouds so that they can be examined visually more clearly. The words in word clouds can be considered keywords belonging to a topic. Showing a word in a more extensive and bolder font means the subject emphasizes it more.

Table 3. Nature of science themes featured in the movie (Seckin-Kapucu, 2016)

Duration		Dialogue	Scientific knowledge
01:22:17,040 01:22:21,238	-->	then the sun's gravitational field has shifted the stars' position	Scientific knowledge is reliable but is subject to change.
01:22:21,280 01:22:24,033	-->	and we have a new theory of gravity.	
00:17:03,120 00:17:07,830	-->	What use is science if it has no practical application?	Scientific knowledge includes logical, mathematical, or experimental inferences.
01:09:39,040 01:09:42,316	-->	As we look at the eclipsed sun through the giant telescope,	
01:09:45,640 01:09:50,111	-->	We'll take photographs of these stars during the five minutes of eclipse,	
01:09:50,160 01:09:52,913	-->	and then compare them to photographs taken of the same stars at night.	
01:23:34,880 01:23:40,352	-->	Einstein says that time is not the same for all of us...	Scientific knowledge is subjective.
01:07:37,160 01:07:39,276	-->	Pick up the tablecloth	Human imagination and creativity have an important role in the acquisition of scientific knowledge.
01:07:43,280 01:07:45,794	-->	Space. The tablecloth is space.	
01:06:19,600 01:06:21,079	-->	Pass, please.	Scientific knowledge is influenced by social and cultural factors at the developmental stage.
01:06:24,840 01:06:28,913	-->	Would you see that this gets to Cambridge?	
01:06:28,960 01:06:31,030	-->	They won't allow me to post it.	
01:06:33,520 01:06:35,954	-->	But you... you are...	
01:06:36,000 01:06:39,470	-->	They would allow you to. Please.	

Experimental Studies on Dataset-1

When the analysis results performed with LDA to extract five topics on Dataset-1 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 2. When the word cloud for five topics in Dataset-1 is reviewed, it is seen that the words "science, max, theory, scientist, Eddington" stand out.

Figure 2. Word clouds of 5 topics extracted by LDA analysis on Dataset-1

When the analysis results performed with LDA to extract seven topics on Dataset-1 (Table 1) are examined, the word cloud related to the resulting word density is presented in Figure 3. As a result of the analysis made to extract seven topics in Dataset-1, the emphasis on the words "science, universe, max, theory, English, feel, Eddington" stands out.

Figure 3. Word clouds of 7 topics extracted by LDA analysis on Dataset-1

When the analysis results performed with LDA to extract nine topics on Dataset-1 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 4. In Dataset-1, the words "science, universe, space, theory, Winnie, max, Albert, question, Eddington" are included in the word density resulting from the analyses made to extract nine topics.

Figure 4. Word clouds of 9 topics extracted by LDA analysis on Dataset-1

When the analysis results performed with LDA to extract 5,7,9 topics on Dataset-1 are examined, it is seen that the most emphasized words are “science, max, theory, eddington”. These words are followed by the word “universe”.

Experimental Studies on Dataset-2

When the analysis results made with LDA to extract five topics on Dataset-2 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 5. As a result of the analysis made to extract five topics in Dataset-2, the words "theory, science, Eddington, Ypres, max" were highlighted.

Figure 5. Word clouds of 5 topics extracted by LDA analysis on Dataset-2

When the analysis results performed with LDA to extract seven topics on Dataset-2 (Table 1) were examined, the resulting word cloud regarding the word density is presented in Figure 6. As a result of the analyses made to

extract seven topics in Dataset-2, the words "Berlin, theory, Eddington, science, scientist, space, max" were seen.

Figure 6. Word clouds of 7 topics extracted by LDA analysis on Dataset-2

When the analysis results performed with LDA to extract nine topics on Dataset-2 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 7. As a result of the analysis made to extract nine topics in Dataset-2, there is a density related to the words "max, Winnie, max, time, berlin, theory, science, Eddington, science."

Figure 7. Word clouds of 9 topics extracted by LDA analysis on Dataset-2

When the analysis results performed with LDA to extract 5,7,9 topics on Dataset-2 are examined, it is seen that the mainly highlighted words are "max, science, theory, eddington".

Experimental Studies on Dataset-3

When the analysis results made with LDA extracting five topics on Dataset-3 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 8. As a result of the analysis made to extract five topics in Dataset-3, a density related to the words "science, Eddington, science, space, theory" was formed.

Figure 8. Word clouds of 5 topics extracted by LDA analysis on Dataset-3

When the analysis results performed with LDA to extract seven topics on Dataset-3 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 9. As a result of the analysis made to extract seven topics in Dataset-3, the words "time, max, science, theory, science, berlin, Eddington" stand out.

Figure 9. Word clouds of 7 topics extracted by LDA analysis on Dataset-3

When the analysis results performed with LDA to extract nine topics on Dataset-3 (Table 1) were examined, the word cloud related to the resulting word density is presented in Figure 10. As a result of the analysis made to extract nine topics in Dataset-3, the emphasis on the words "max, science, theory, Eddington, science, scientist, theory, space, berlin" stands out.

Figure 10. Word clouds of 9 topics extracted by LDA analysis on Dataset-3

When the analysis results performed with LDA to extract 5,7,9 topics on Dataset-3 are examined, it is seen that the most emphasized word is "science," and this word is followed by the words "theory, Eddington". When the prominent eliminations in the three data sets are examined, it is seen that the majority of the emphasis is on the words "science, theory, Eddington, and max."

Discussion

When the word clouds that emerged because of the analysis to extract 5-7-9 topics in three datasets made on the Einstein and Eddington documentary film are examined, it is seen that words such as science, scientist, theory, universe, Eddington, think, speed, answer stand out. In addition to that, the LDA model appears as a method used to identify NOS themes by some other studies. In the study, published by Wang et al. in 2023, it was found that other issues other than the NOS were also identified. The 12 topics obtained in the relevant study were associated with eight NOS themes due to manual reviews. The research done by Wang et al. (2022) is similar to the current study. LDA topic modeling not only considers sentences containing terms, keywords, and phrases that contain a view of NOS but also analyzes the entire perspectives of participants (Wang et al., 2023). The words and topics that emerge with LDA are insufficient to determine the nature of science themes. Because the same terms, expressions, or keywords have different meanings in different contexts (Jaeger et al., 2019). If we consider the word “science” without context, the meaning is the same, but the meaning of the word “science” is different in different contexts. The use of the same word in different subjects provides a context for the word, and the meaning expressed is different (Wang et al., 2023). The words that appear in the analysis are words used in the nature of science themes but also words in the content of the documentary film. For this reason, it seems difficult to decide on the nature of science themes based on words.

Another result obtained in this study is that although the data size is the same, the results obtained with Dataset-1, which is handled in 30-second sections, reveal more sensitive subject sections in the context of keywords compared to Dataset-2 and Dataset-3, which have 1-minute and 3-minutes sections. When the studies conducted

with LDA are examined, large data sets are used (Gurcan & Çağıltay, 2019; Na, 2024; Yuan et al., 2020). Since the number of words to be studied in this study was limited, the data were divided according to certain time intervals, and datasets containing the same data but constructed in a different structure were created. The result of this study's division of data sets differs from other studies.

Limitations

In this section, we will discuss the limitations and threats of our work. Although we have carefully established the research infrastructure, potential threats to the validity of our results and conclusions may still exist. In this section, we consider these threats and their possible impact on the results of our study.

LDA detects unknown issues within the documents given due to its structure and the results it proposes. In this research, the problems related to NOS are tried to be determined, and other issues and scopes may be gathered. This is among the expected threats.

LDA does not deal directly with the titles of the topics. It produces word sets that express a subject in a way that emphasizes the one that expresses it the most. In this case, the necessity of matching NOS topics with human intervention emerges. This situation is a common limitation in almost all of these similar studies.

Internal Validity. Although NOS topics are clear and limited, these topics may not have been mentioned in the documentary contents discussed in the research. In this case, choosing the number of topics 5-7-9 may have prevented more accurate determinations.

External Validity. When checking the results obtained with LDA, the results obtained from a previous study with subjective approaches were taken as references. Although the subjective analyses were carefully conducted, imprecision may be possible.

Conclusion

This study aims to examine the NOS themes in the documentary film "Einstein and Eddington" according to the LDA Model and to extract the NOS themes in the documentary film. When the word clouds that emerged as a result of the analysis to extract 5-7-9 topics in three datasets made on the Einstein and Eddington documentary film are examined, it is seen that words such as science, scientist, theory, universe, Eddington, think, speed, answer stand out. However, when these words are reviewed to evoke the NOS themes, it is pretty challenging to identify the NOS themes in the documentary film under study. Topic extraction models enable topics other than the NOS themes in this documentary film to be revealed and even come to the fore more dominantly.

Another result obtained in this study is that although the data size is the same, the results obtained with Dataset-1, which is handled in 30-second sections, reveal more sensitive subject sections in the context of keywords compared to Dataset-2 and Dataset-3, which have 1-minute and 3-minutes sections. Although the total number of words is the same, more divisions may cause words to play a more distinctive role in the frequency of their inclusion in the sections. An ideal time interval may have formed more appropriate and unique phrases in this case.

On the other hand, it is predicted that topic extraction algorithms such as LDA can produce more meaningful word clouds on much larger datasets. Since large datasets are needed for artificial learning algorithms to work effectively, it can be said that it is not enough for the algorithm to produce successful results. Since topic extraction methods such as LDA produce results by extracting frequencies on the data without training, the detected phrases also cover topics other than the NOS themes.

In line with the increasing importance of scientific literacy, it is considered that the NOS themes is quite challenging to understand and contains a lot of subjectivity; artificial intelligence-supported systems need to identify these complex topics and make them easier to understand. The results obtained by this study are considered one of the first stage studies on detecting the NOS themes with artificial intelligence models. In this respect, it will guide artificial intelligence-supported studies on the NOS.

Recommendations

For this reason, it is essential to use artificial learning models trained with the information in the datasets in future studies to focus directly on the NOS themes and make more accurate determinations. It is thought that very successful results can be obtained from analyzing the transcripts of many documentary films that will discuss the NOS with artificial learning models (deep learning, machine learning, artificial neural networks, etc.). For computer-assisted detection of the NOS themes, it is recommended to study artificial learning models trained with data rather than topic extraction algorithms such as LDA. In addition, data obtained from studies on the nature of science or science education can be analyzed using various artificial learning models. Analyses made using different artificial learning models can be compared with each other or with analyses made with computer programs.

References

Abd-El-Khalick, F., Bell, R. L., & Lederman, N. G. (1998). The nature of science and instructional practice: Making the unnatural natural. *Science Education*, *82*(4), 417-436.

Adams Becker, S., Freeman, A., Giesinger Hall, C., Cummins, M., & Yuhnke, B. (2016). *NMC/CoSN horizon report: 2016 K-12 edition*. The New Media Consortium

Allen, C., & Murdock, J. (2020). LDA Topic Modeling: Contexts for the History & Philosophy of Science. Forthcoming in Ramsey, G., De Block, A.(Eds.) *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*. Pittsburgh University Press; Pittsburgh.

American Association for the Advancement of Science. (1989). *Science for all Americans.* New York: Oxford University Press.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy.* New York: Oxford University Press.

Başkalyoncu, H. (2017). *The effects of documentary films of historical science in the teaching of the nature of science and the particulate nature of matter* [Unpublished master's dissertation]. University of Abant İzzet Baysal.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27–40.

Calvo, H., Hernández-Castañeda, Á., & García-Flores, J. (2018). Author identification using latent Dirichlet Allocation. In: Gelbukh, A. (eds) *Computational Linguistics and Intelligent Text Processing*. CICLing 2017. Lecture Notes in Computer Science(), vol 10762. Springer, Cham. https://doi.org/10.1007/978-3-319-77116-8_22

Celik, U. (2018). BIST30 verilerinin R programlama dili ile zaman serisi analizi - Time series analysis of BIST30 data with R programming language. II. *International Symposium on Economics, Finance and Econometrics (ISEFE,18),* Türkiye

Dagher, Z, R. & Boujaoude, S. (2005). Students' perceptions of the nature of evolutionary theory. *Science Education, 89(3)*, 378-391.

Dark, M. (2005). *Using science fiction movies in introductory physics*. *The Physics Teacher*, *43*, 463-465. https://doi.org/10.1119/1.2060648

Efthimiou, C. J., & Llewellyn, R. A. (2006). Avatars of Hollywood in physical science. *The Physics Teacher*, *44*, 28–33. https://doi.org/10.1119/1.2150756

Efthimiou, C. J., & Llewellyn, R. A. (2007). Cinema, fermi problems, and general education. *Physics Education*, *42*, 253–261. https://doi.org/10.1088/0031-9120/42/3/003

Güven, Z. A., Erdoğan, R. T., Diri, B., & Çakaloğlu, T.(2018). *n-seviyeli gizli Dirichlet ayırımı ile Türkçe tivit duygularının sınıflandırılması Classification of Turkish tweet emotions by n-stage* Latent Dirichlet Allocation.

Gurcan, F., & Cagiltay, N. E. (2019). Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE access*, *7*, 82541-82552.

Hand, B., Prain, V., Lawrence, C. & Yore, L. D. (1999). A writing in science framework designed to improve science literacy. *International Journal of Science Education*, *10,* 1021–1036.

Hodson, D. (2008). *Towards scientific literacy, a teachers' guide to the history, philosophy and sociology of science*. Sense Publishers.

Hornik, K. (2012). The comprehensive R archive network. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(4), 394–398.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299–314.

Irez, S. (2006). Are we prepared?: An assessment of preservice science teacher educators' beliefs about nature of science. *Science Education, 90(6),* 1113–1143.

Jaeger, S. R., Roigard, C. M., Jin, D., Vidal, L., & Ares, G. (2019). Valence, arousal and sentiment meanings of 33 facial emoji: Insights for the use of emoji in consumer research. *Food research international* (Ottawa, Ont.),119, 895–907.https://doi.org/10.1016/j.foodres.2018.10.074

Jiang, H., Qiang, M., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. *Renewable & Sustainable Energy Reviews*, *57*, 226–237. https://doi.org/10.1016/j.rser.2015.12.194

Johnson, L., Adams Becker, S., Estrada, V., & Freeman, A. (2015). *NMC horizon report: 2015 higher education edition*. The New Media Consortium.

Kartal, Y. (2017). *Topic modelling of TOJDE journal with LDA* [Unpublished master's dissertation]. Anadolu University.

Kaya, A., & Gülbandılar, E. (2022). Comparison of subject modelling methods. *Eskişehir Turkish World Application and Research Centre Informatics Journal*, *3*(2), 46-53.

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching, 29*(4), 331-359.

Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 831–880). Mahwah, NJ: Lawrence Erlbaum Associates.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching*, *39*(6), 497-521.

Li, W., & McCallum, A. (2006, June). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 577–584).

Lin, C., & He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 375–384).

McComas, M. R. (2003). A textbook case of the nature of science: laws and theories in the science of biology. *International Journal of Science and Mathematics Education,*1, 141–155.

McComas, W. F. (1998). The principal elements of the nature of science: Dispelling the myths of science. In W. F. McComas (Ed.) *The nature of science in science education: Rationales and strategies*. (pp. 53–70). Kluwer Academic Publishers.

Na, J. C., Kim, E. J., & Kim, J. Y. (2024). Unveiling metaverse social trends: Analysing big data regarding online sports news with LDA-based topic modelling. *Revista de Psicología del Deporte (Journal of Sport Psychology)*, *33*(1), 115-125.

National Research Council (1996). *National Science Education Standards.* National Academy Press: Washington, DC.

National Science Teacher Association (1982). *Science-Technology-Society: Science education for the 1980s* .Washington, DC: NSTA.

Navakanesh, B., Shah, A. A., & Prasanna, M. V. (2019). Earthquake education through the use of documentary movies. *Frontiers in Earth Science*, *7*, 42. https://doi.org/10.3389/feart.2019.00042

Özdemir, A. F., Yıldıztepe, E., & Binar, M. (2010). İstatistiksel yazılım geliştirme ortamı: R. *XII. Akademik Bilişim Konferansı,* 10-12.

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

Piliouras, P., Siakas, S., & Seroglou, F. (2011). Pupils produce their own narratives inspired by the history of science: Animation movies concerning the geocentric heliocentric debate. *Science & Education*, 20(7-8), 761-795. https://link.springer.com/article/10.1007/s11191-010-9321-4

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012, May). EmpaTweet: Annotating and detecting emotions on Twitter. In *Lrec* (Vol. 12, No. 12, pp. 3806-3813).

Ryan, A. G., & Aikenhead, G. S. (1992). Students' preconceptions about the epistemology of science. *Science Education*, *76*(6), 559–580.

Seckin-Kapucu, M. (2023). Analyzing the documentary film" Marie Curie" in terms of the nature of science themes. *The Eurasia Proceedings of Educational and Social Sciences*, *31*, 17-25.

Seçkin-Kapucu, M. (2016). An examination of the documentary film" Einstein and Eddington" in terms of nature of science themes, philosophical movements, and concepts. *International Journal of Progressive Education*, *12*(2), 34-46. https://ijpe.inased.org/files/2/manuscript/manuscript_3/ijpe-3-manuscript-224451.pdf

Seçkin-Kapucu, M., Cakmakci, G., & Aydogdu, C. (2015). The influence of documentary films on 8th grade students' views about nature of science. *Educational Sciences: Theory & Practice*, *15*(3), 797-808, https://doi.org/10.12738/estp.2015.3.2186

Shen, B. S. (1975). Views: Science Literacy: Public understanding of science is becoming vitally needed in developing and industrialized countries alike. *American Scientist, 63*(3), 265–268.

Smith, M. U., Lederman, N. G., Bell, R. L., McComas, W. F., & Clough, M. P. (1997). How great is the disagreement about the nature of science: A response to Alters. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *34*(10), 1101–1103.

Smith, U. M., & Scharmann, L. C. (1999). Defining versus describing the nature of science: A pragmatic analysis for classroom teachers and science educators. *Science Education*, *83*(4), 493–509. https://doi.org/10.1002/(SICI)1098-237X(199907)83:4<493::AID-SCE6>3.0.CO;2-U

Tur, G., Celikyilmaz, A., & Hakkani-Tür, D. (2013, May). Latent semantic modeling for slot filling in conversational understanding. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8307-8311). IEEE.

Wang, M., Gao, S., Gui, W., Ye, J., & Mi, S. (2023). Investigation of pre-service teachers' conceptions of the nature of science based on the LDA model. *Science & Education*, *32*(3), 589-615.

Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *29*, 142-156.

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.

Yıldırım, C. (2010). *Bilim felsefesi*. Remzi Kitapevi.

Yıldırım, E. G., Köklükaya, A., & Selvi, M. (2015). Öğretim materyali olarak 3-Idiot filmi ile öğretmen adaylarının günlük hayatta fenin kullanımı ve eğitimde aile rolü üzerine görüşlerinin belirlenmesi. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, *5*(2), 94-105. https://dergipark.org

Yılmaz, M. (2018). Filmlerin öğretim materyali olarak kullanılması ve biyoloji eğitimindeki yansımaları. *İnformal Ortamlarda Araştırmalar Dergisi*, *3*(2), 24-37. https://dergipark.org.tr/en/pub/jrinen/issue/42184/427249

Yin, B., & Yuan, C. H. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. *Education and Information Technologies*, *27*(9), 12689-12712.

Yuan, L., Bin, J., Wei, Y., Huang, F., Hu, X., & Tan, M. (2020). Big data aspect-based opinion mining using the SLDA and HME-LDA models. *Wireless Communications and Mobile Computing*, *2020*(1), 8869385.

Zhao, L., Zhao, Q., & Wang, Y. (2020, October). Research on Chinese movie reviews based on latent Dirichlet Allocation Topic Model. In *2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 46–49). IEEE.

Zhou, Z. P., Zhou, X. N., & Qian, L. F. (2021). Online public opinion analysis on infrastructure megaprojects: Toward an analytical framework. *Journal of Management in Engineering*, *37*(1). https://doi.org/10.1061/(ASCE)Me.1943-5479.0000874

Author(s) Information

Yusuf Kartal
Eskisehir Osmangazi University
Computer Engineering Dept. Eskişehir/Türkiye
Contact e-mail: ykartal@ogu.edu.tr
ORCID iD: https://orcid.org/0000-0002-0402-1701

Munise Seckin-Kapucu
Eskisehir Osmangazi University
Faculty of Education, Eskisehir/ Türkiye
ORCID iD: https://orcid.org/0000-0002-9202-2703