Advanced Engineering Days

aed.mersin.edu.tr

The implementation of Latent Dirichlet Allocation (LDA) model on IEEE Xplore dataset to find the impact of artificial intelligence in education sector

Olgerta Idrizi *1, **Alfons Harizaj** 2, **Miranda Harizaj** 3

1 *Mediterranean University of Albania, Faculty of Informatics, Albania, idriziolgerta@gmail.com*
2 *Canadian Institute of Technology, Faculty of Engineering, Albania, alfons.harizaj@cit.edu.al*
3 *Polytechnic University of Tirana, Faculty of Electrical Engineering, Albania, miranda.harizaj@fie.edu.al*

Cite this study: Idrizi, O., Harizaj, A. &, Harizaj, M. (2023). The implementation of Latent Dirichlet Allocation (LDA) model on IEEE Xplore dataset to find the impact of artificial intelligence in education sector. Advanced Engineering Days, 7, 145-147

Keywords
Latent Dirichlet Allocation
LDA
Artificial Intelligence
Education Sector
Machine Learning

Abstract

The field of education holds immense significance, calling for a reevaluation of learning methods and approaches. Particularly in recent years, there has been a growing inclination within higher education to incorporate emerging technologies and artificial intelligence (AI) in order to enrich the learning process. This study aims to analyze the Latent Dirichlet Allocation LDA, as a probabilistic Bayesian model designed for analyzing collections of discrete data, such as text corpora. LDA employs a three-level hierarchical Bayesian model, where each item in the collection is represented as a finite mixture derived from a set of underlying topics. These topics, in turn, are modeled as an infinite mixture based on a set of topic probabilities. In the realm of text modeling, the topic probabilities offer a transparent representation of a document's content. The descriptive analyze work on data collection from IEEE Xplore from 2011 to 2022 years.

Introduction

Artificial Intelligence (AI) in Education involves the integration of AI techniques into traditional learning methods, with the aim of automating or replicating existing educational practices. However, much of the focus has been on replacing or diminishing the role of teachers, rather than assisting them in improving their teaching effectiveness. While this approach may be beneficial in areas with limited access to teachers, it fails to recognize the unique skills and experiences that teachers bring, as well as the importance of social learning and guidance for learners.

Instead of simply automating computer-based instruction, AI has the potential to expand the possibilities of teaching and learning that are otherwise difficult to achieve. It can challenge existing pedagogical approaches or assist teachers in enhancing their effectiveness. AI can support collaborative learning by facilitating AI-driven monitoring of student forums, enable AI-powered continuous assessment, provide AI learning companions for students, and offer AI teaching assistants for teachers. These applications of AI have the potential to revolutionize the educational landscape [1].

Furthermore, AI in Education can serve as a valuable research tool in the field of learning sciences, advancing our understanding of the learning process. By exploring the possibilities and limitations of AIED, we can uncover new insights and contribute to the improvement of educational practices.

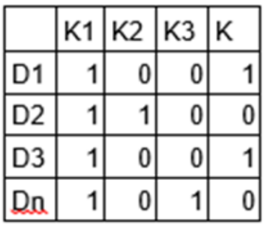
In conclusion, this paper examines the challenge of modeling text corpora and other collections of discrete data from IEEE Xplore dataset. The objective is to discover concise representations of the items in a collection that allow for efficient processing of large datasets, while retaining the crucial statistical relationships that are valuable for tasks such as classification, novelty detection, summarization, and assessing similarity and relevance [2].

Latent Dirichlet Allocation LDA

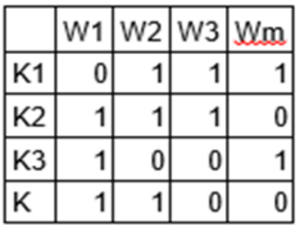
Latent Dirichlet Allocation (LDA) has gained significant recognition as a prominent method in the field of topic modeling. This approach is particularly effective in analyzing collections of discrete data, such as text corpora. LDA operates as a generative probabilistic framework, employing a three-level hierarchical Bayesian model. Each item in the collection is represented as a mixture of predefined topics, while the topics themselves are modeled as an infinite mixture based on topic probabilities. When applied to text modeling, LDA provides a comprehensive representation of a document's content, allowing for insightful analysis of its underlying topics and themes. Below are three steps explaining LDA model [3].

Step-1

Latent Dirichlet Allocation (LDA) transforms a Document-Term Matrix into two matrices, M1 and M2, which are of lower dimensions. Matrix M1 represents the document-topics relationship, while matrix M2 represents the topic-terms relationship. Matrix M1 has dimensions (N, K), where N corresponds to the number of documents and K represents the number of topics. On the other hand, matrix M2 has dimensions (K, M), where K is the number of topics and M signifies the size of the vocabulary. We have to improve these distributions, which is the main goal of LDA [4].



	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0



	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

Step-2

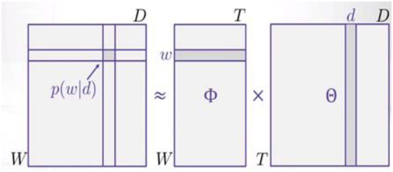
During this step, we go through each word "w" in every document "d" and aim to modify the current assignment of topics for the word. We assign a new topic "k" to the word "w" based on a probability P, which is determined by multiplying two probabilities, p1 and p2. To calculate these probabilities for each topic.

Step-3

During this step, the model operates under the assumption that all the word-topic assignments, except for the current word, are accurate. The model calculates the probability that a specific topic "t" generated the word "w". Based on this probability, it adjusts the assignment of the current word to a new topic. This adjustment is made iteratively, and over time, the model reaches a steady-state where the distributions of document topics and topic terms become reasonably accurate. This steady-state is considered the convergence point for Latent Dirichlet Allocation (LDA). We derived that P(w|d) is equal to: [2].

$$\sum_{t=1}^{T} p(w|t)\, p(t|d)$$

The above thing can be also represented in the form of a matrix (shown below):



By examining the provided diagram, we can draw a parallel between Latent Dirichlet Allocation (LDA) and matrix factorization or singular value decomposition (SVD). In both cases, we aim to decompose the probability distribution matrix of words in documents into two matrices: one representing the distribution of topics in a document and the other representing the distribution of words in a topic. This decomposition allows us to uncover the latent structure and relationships within the data.

Descriptive analyze

Based on the analysis of the IEEE Xplore dataset, the generated figure clearly illustrates a notable upward trend in the publication of papers focusing on the intersection of AI and Education from 2011 to 2022. Over the span of ten years, there has been a significant increase, particularly in the recent years, with a substantial portion of these papers being presented at conferences.

In the second figure, it is observed that the proportion of papers containing the keyword "Education" within the subset of papers that also include the keyword "AI" remains relatively consistent from 2011 to 2022. This finding is unexpected since one would anticipate an increase in this percentage over time. However, the reason behind this observation can be attributed to the overall rise in the total number of papers published within the dataset.

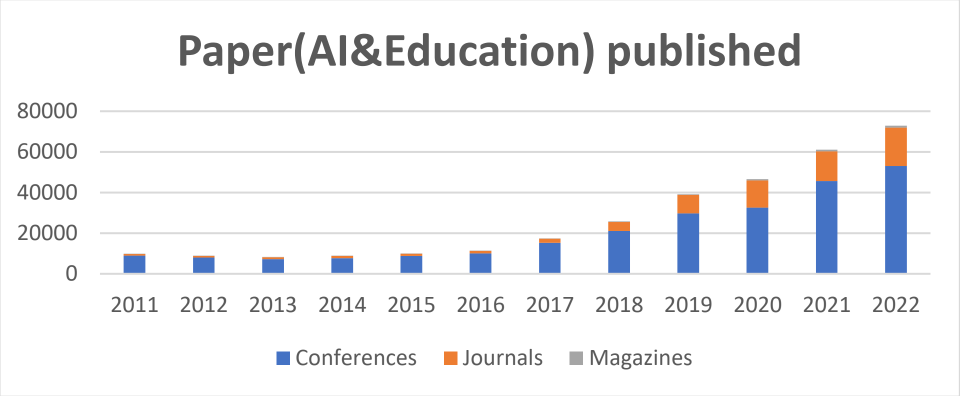


Figure 1. Papers in IEEE Xplore in the last years with key words “AI” and “Machine Learning” in Education [5]

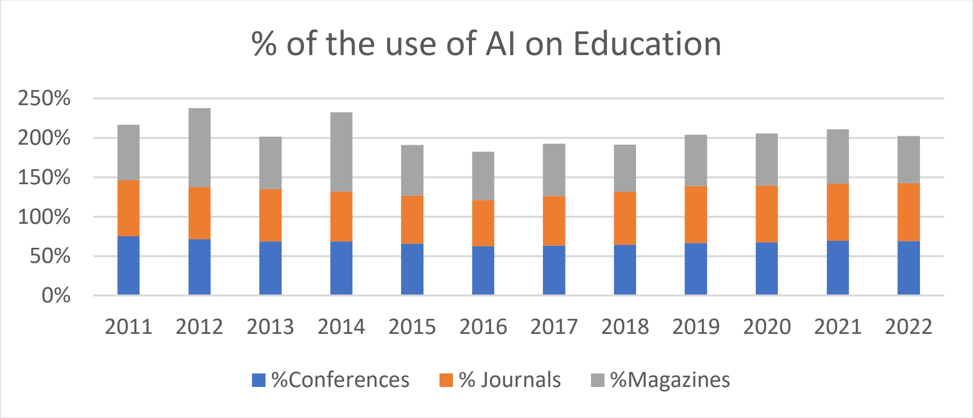


Figure 2. The percentage of Papers in IEEE Xplore in the last years with key words “Education” in papers with key words “AI” and “Machine Learning” [5]

Results and Conclusion

The application of AI algorithms and systems in education has garnered increasing interest over the years. Figure 1 illustrates the growing number of research papers published on the topics of "AI" and "Education" since 20101 based on data from IEEE Xplore. Throughout the period from 2011 to 2022, there is a notable consistency in the proportion of papers that incorporate the keyword "Education" within the subset of papers that also include the keyword "AI." As the field of education undergoes continuous development, scholars are actively investigating the application of cutting-edge AI techniques like deep learning and data mining. Their aim is to tackle intricate challenges and tailor teaching approaches to suit the unique needs of individual students.

References

1. Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264-75278.
2. Goyal. C. (2021). Step by Step Guide to Master NLP – Topic Modelling using LDA (Matrix Factorization Approach)
3. Paek, S., & Kim, N. (2021). Analysis of worldwide research trends on the impact of artificial intelligence in education. Sustainability, 13(14), 7941.
4. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
5. Harizaj, M., Idrizi, O., & Harizaj, A. (2023). Machine learning algorithms for predicting life expectancy. Advanced Engineering Days (AED), 6, 132-134.