Extract Concept using Subtitles in MOOC

Aarika Kawtar*, Habib Benlahmar, Mohamed Amine Naji, Elfilali Sanaa, Zouheir Banou
Laboratory of Information Technology and Modeling, Hassan II University of Casablanca, Faculty of Sciences
Casablanca, Morocco

***Abstract*—Massive open online courses (MOOCs) are a variety of courses offered through the online mode, paid or unpaid and has evolved as an excellent learning resource for students. The structure of the course design is mainly linear where there are a few video lectures provided by either professors of several universities, or people with expertise in the particular subject. They are usually graded on a weekly basis through quizzes or peer-graded assignments. The objective of this paper is to extract the concepts taught in the videos from the subtitles, which could later be used to enhance recommendations of the learners using their clickstream data. The teachers could also use this to see the demand for their courses. Evaluate two keyword extraction methods, which are BERT and LDA using different Coursera courses. The experimental results show that BERT outperforms LDA in terms of Coherence.**

Keywords—LDA; BERT; topic coherence; overlap coefficient

I. Introduction

One of the most profitable and in demand businesses in today’s world are those of Massive open online courses (MOOC). Not only do they offer a vast range of lectures on almost all the topics be it the medical field, or some complex lessons on coding, but can also be easily accessible by everyone sitting at home [1]. These platforms have attracted a large number of people, which sums up to nearly 10 million participants from all over the world. Coursera, Udemy and EdX are some of the classic examples of MOOC [2]. The format of these platforms is similar, where professors or trained people share video lectures covering a particular topic. They use different methods of teaching like using powerpoint presentations, whiteboard or even the electronic boards. Some platforms invest a lot in making their videos interesting and visually appealing, hence they incorporate graphics and colorful animations. This helps in drawing more attention from the learners, especially the younger crowds. The lectures are usually grouped into few modules and a set of modules makes up a course. The module system helps the learner keep track of their progress and also have a better understanding of the pre requisites. The teachers find the module systems easier as it is easier to set assignments and other assessment related work. Each module usually runs for a week, however, it totally depends on the viewers’ interests. Weekly deadlines are set, which are flexible. This means that an ideal schedule is provided, which if followed thoroughly can benefit an average learner. However, it is their choice ultimately on how much time they want to spend on it, it could be earlier than the target date of even later. At the end of each module, there is an assessment held. There are several ways in which one is assessed to see how much of the course they have grasped. Some of these assessment techniques are quizzes, projects and peer-graded assignments. Few courses even have cutoffs to be cleared at the end of module assessment. Failure to complete this successfully would not permit the learner to proceed to the next module or it might not consider the module as complete [3]. Upon course completion, the learner receives a certificate of completion from the institute offering it and it can be considered as a legitimate proof of knowledge acquired, and can be updated in resumes and professional profiles. Courses that have strict assessments do not provide the certificates until all the quizzes have been cleared with the minimum required cutoff and all the peer graded assignments have been checked by the required number of co learners. The legitimacy of MOOC has gone so far that nowadays, universities offer these courses as electives as proper curriculum courses with college credits awarded on their completion. The college provides these courses and has their own assessment methods, however, the students have to complete these courses through the platform in order to receive the assigned number of credits. These courses can be free, but mostly they have to be purchased. Another alternative provided by MOOC is that some courses can be audited for free but do not provide completion certificates hence the purpose is solely for acquiring knowledge.

MOOC provides a form of social learning where interactions constantly take place between learners and the teachers. It paves way for mass learning and personalised comprehension. Even though there is no face-to-face communication taking place, these platforms have been successfully been able to break the barriers of any type of communication hindered otherwise [4]. There are different ways one can engage themselves with the platforms. Learners and teachers can both participate in forum and discussions, helping fellow learners and students. Some even start taking lectures of their own. Others work on in video editing, as mentioned earlier, adding good graphical depictions of what is being explained or colourful animations. There is a lot work that has to be dealt in the back end of the sites or apps belonging to these platforms. A large group of people also contributes by providing constructive feedback and suggests improvements. These are constantly monitored and taken note of in order to improve the user interface of their platforms and attract more learners to purchase their products. These learning methods are completely different from the physical mode of learning and open a wide door of new opportunities to explore [5]. Hence, we can say that the most important factor which determines the success of these MOOCs is the engagement of the students, however not a lot of research has been carried out on how the student engagement affects the platforms. All MOOC platforms primarily run on how much they have been used and a decline in student engagement can give a massive blow to these businesses. It is of utmost importance that the

*Corresponding Author

engagement is always constantly monitored and changes being continuously implemented in order to keep them high [6]. The discussion forums play a vital role in checking engagement, along with website visits, registrations, clicks etc. However, it is not an easy task to keep track of the engagement as there are so many parameters that have to be taken into consideration while doing the analysis. Some of them are course enrollments, course completion, discussion forums, etc. [7].

These courses some with their own set of disadvantages. Though they attract a large number of student registrations, recent studies have shown that only a small fraction of these students complete their courses [8]. According to statistics provided by Coursera, almost 75% of the courses enrolled by students have not been completed [9]. Another problem is that these platforms do not come with keyphrases and it is going to be a laborious task to identify them manually and will take up a lot of time. This means that one cannot search for courses based on particular topics. There are a variety of topics mentioned in each video, but there is no way of keeping track of these. It is important to do so as it can help recommending better courses to those who show interest in topics. Topic based searches can be made than course-based searches and it will be easier for the learner to choose their apt course based on how much do the topics cover in the course line up with their topics of interest.

Keyphrases are important and significant expressions consisting a collection of words. They give us the contents of the data, or even sometimes summarize it [10]. There have been several algorithms developed to extract keywords from scripts, notes etc. These are used in data mining like clustering of documents, providing recommendations and formulation of queries [11], [12]. Bidirectional Encoder Representations from Transformers (BERT) is one of the models that can be used for keyphrase extraction. This model is used to make sequential recommendations based on past data. The distinctive feature of this method is that it can incorporate context from both sides, unlike other sequential predictors, which only do it from left to right [13]. Latent Dirichlet allocation (LDA) model is a probabilistic modeling algorithm. It is commonly used to identify the topics in a collection of texts. It is usually used in image retrieval and face recognition technologies [14], [15].

Instructors face problems in analyzing each student's level of understanding in order to improve the quality of courses or to provide referral systems. Although the number of students enrolling in courses has increased, very few of them actually complete the course. Therefore, it is necessary to track learner journey data to know what interests them. The goal of this paper is to extract the concepts taught in the videos from the subtitles, which could then be used to improve the learners' recommendations using their path data. Instructors could also use this to learn about the demand for their courses.

In this paper, we have attempted to extract concepts from the subtitles of video lectures of courses offered by Coursera using BERT and LDA models for key phrase extraction. A comparison is made between the results obtained by both.

The paper is organized as follows. In Section 2, we review related work on concept extraction.

Section 3 is devoted to the context of our experimental study, detailing the dataset collected from the Coursera MOOC videos and the models (LDA and BERT) that we will use for this study.

In Section 4, we show our proposed algorithms for concept extraction from the sub-titling of the experimental results which show a better concept extraction. In section 5 we end with a conclusion that shows the results of our work.

II. Related Work

In our study, we have tried to automatically extract keyphrases from the subtitles of the videos. In general, there are two ways in which these extractions are carried out [16]. The first approach is supervised, where there is binary segregation of each word into either keyphrase or not a keyphrase [17]. The second approach is unsupervised. In this approach, the words are ranked based on what the algorithm asks it to do, for example probability of occurrence, or even usage in the course. Some commonly used machine learning algorithms are Naïve Bayes and support vector machines [18].

Yi-fang et al developed an algorithm called KIP algorithm. In this algorithm the extracted words were first examined and scored on the basis of three factors. The first factor was their frequency in the text. This means they checked how frequently the word occurred in the text. Second parameter considered was their specificity. This means there is a check on how specific or unique the words are to the course provided. This information is also gathered by checking on the neighborhood data. Last parameter taken into consideration is its contents, as in the words that are related to the examined word. The words are arranged in order of their scores. The words that obtain high scores are later categorized as keyphrases [12]. Another similar type of work can be seen in Xiaojun et all's paper used information from the neighborhood documents to get more data and then this data was graphically represented along with the data of the document where keywords need to be extracted. These data were compared and the keyphrase were extracted accordingly [10]. A very similar study to ours was found in the works of Raga et al. They used this method to navigate to the exact part of the video, or access a video segment by just searching for the keyword. They considered factors like statistical and visual features while implementing the algorithm [19]. A model called Text Rank was developed by Rada and Paul where they took a graphical approach to rank the words [20]. The TPR (Topical Page Rank) approach is another one proposed by Liu et al where first the segregation occurs based on various topics and then the TPR algorithm is individually run on each one [18]. Some other algorithms were developed based on using deep learning [21] , frequency of occurrence of words [22], word embedding vectors and graphical ranking [23].

Our study draws inspiration from all these works, but still manages to stand apart as we aim on increasing engagement by giving personalised recommendations to learners based on their search history or clickstream data. To ensure this, keyphrases have been extracted from the subtitles using two algorithms and the best of these two on experiments could be used to develop better recommendation systems. Elaborating on that, these keyphrases can be used to match with the learner's interests thus giving better course recommendations. The teachers can also benefit from this study. The clickstream data allows teachers to know which topics are more in demand and will encourage them to record lectures covering those topics. This will help the algorithm detect their courses and recommend it to the learners. They can also gauge the learner engagement and see which part of their courses attract more attention.

III. Methodology

A. Datasets

The study will use the dataset called "MOOC DATA". This dataset has been derived from the subtitles of the course videos from Coursera platform. All the words are split up into individual components and these could further be sent into algorithms for keyphrase extraction. This dataset consists of a total of four folders named:

- "CSEN" – Computer Science in English
- "CSZH"- Computer Science in Chinese
- "EcoEN"- Economy in English
- "EcoZH"- Economy in Chinese

The statistics of the four datasets are listed in Table I, where #courses, #videos, are the total number of courses, videos, in each dataset.

TABLE I. Dataset Statistics

Dataset	***Domain***	***Language***	***#courses***	***#video***
CSEN	Computer Science	English	18	2,849
EcoEN	Economics	English	5	381
CSZH	Computer Science	Chinese	8	690
EcoZH	Economics	Chinese	8	455

However, for the sake of better understanding and better research, only the "CSEN" folder was used for the study. This folder contained two JSON files, one of these files was called "candidates" and the other one was called "captions". Again, since the aim of our study only deals with the subtitles of the videos, only the caption file was utilized. This table contained the video captions of 18 computer courses; the size of this file is 216 MB. Table II shows how the subtitles were sliced and stored.

The first column is usually ignored as it is the serial numbers. The second column is the Course ID. This is a unique code, which is used to identify the particular course. For the course we considered (Computer Science with English), the course id is 1. The column next to this is the text. It consists of the script of the subtitles and is called transcript. This script is so precise that it also has details like parts of the video where there is music. That’s what makes the process of keyword extraction challenging. The music is used way too many times; the system might mistake it to be a keyphrase while we know it is just the background music being referred. Hence it is important that all these unwanted parameters are taken care of at the initial stages and the algorithm is not affected due to them. The column next to it is the tagged column. Here, we see that every word has been sliced up (including the music). The 3 gram model is used to carry out this process. For example, the first row shows that the text has been separated into tags like.

“MUSIC”, “Today”, “we”, “re” and so on. Again, on running models, the music words should not be considered for keyphrase extraction. The last column is video id which is the unique number given to the video in a course, and is used to refer to a particular course. The course is the same while the videos from which the words were extracted from were different. Our study uses the first 5 videos from the course with course id 1. This dataset will run through two models and a comparison will be drawn regarding which is the better one to consider for keyphrase extraction.

B. Pre-Processing

Pre-processing is the process where the raw data received is converted into a form that is comprehensible and useful. It is extremely crucial to ensure that data pre-processing has been done before carrying out any analytical task [24] [25]. This helps in having a dataset of good quality. Process used to split the text or segmenting a text to words, meaningful parts or phrases is called tokenization. In this process, punctuations, whitespaces and other non- alphanumeric characters are not considered, all characters are converted to lowercase and stopwords (conjunctions, articles, etc.) are removed [26].

Before proceeding, there is another concept that needs to be looked upon, which is an n-gram. An n-gram (or Q gram) is basically a sliced part of a longer string consisting of n characters. They are usually obtained from a sample text or some form of speech [27]. It could be words, phrases, letters sometimes even syllables. It is a very efficient means of implementation. On conversion in n-grams the string gets embedded into a vector and is further compared with other data of similar type. Its consistency and distribution can be measured too [28]. An n-gram model is a probabilistic language model, where it is used to make predictions of the items succeeding it in the form of a sequence known as an (n-1)- order Markov model. These models find their extensive usage in computational linguistics, communication theory and data compression. There are two major advantages of using thesen-gram models and algorithms. One of them is simplicity, the model is comparatively simpler to operate and execute than its other counterparts. Secondly, its scalability is a boon. At higher n values, this model is able to store more contexts with a space- time tradeoff which has been understood well. This allows the smaller experiments to efficiently expand.

In our study data has been obtained beforehand from the dataset in order to run it with various algorithms. The words from the video subtitles have been sliced out into different words. Each of these go through the algorithm to obtain results on whether it is a keyword or not. That is decided based on other data like how frequently the word is used or its significance in the text. This process will help identify the key topics covered in the course. The data was retained from the dataset, however, unnecessary information like the tagged column and stopwords were eliminated altogether and n-grams were generated.

C. Models used

1) BERT: Bidirectional Encoder Representations from Transformers (commonly known as BERT) is a machine learning model that is used for language representation [28]. These models are pre- trained and they force the model to study the semantic data in between and withing the sentences. Unlike other similar models which only function from left to right, BERT works from both directions i.e. it is bidirectional, just as its name says [29]. This algorithm takes the final hidden state of the first token and uses it to represent the entire sequence for tasks which require classification of texts. When BERT is incorporating with another output layer, there is an advantage of minimal number of parameters being necessary to be learnt form scratch [30]. There is a particular format that any input data needs to fulfill if it has to undergo the BERT model. A special token which consists of the special classification embedding called [CLS] is put prior to every sentence to fulfill this criterion. Another special token that is used is the [SEP]. It is placed at the end of each and every sentence in order to make a clear separation between the segments [26]. BERT also relieves the problem of masked language model (MLM), where it randomly covers some of the input and expects the algorithm to predict it based on the date of the surrounding words. The next sentence prediction (NSP) is also used. Fine – tuning techniques are of various kinds based on how much of the architecture needs to be trained [31]. Basically, it is a sequential predictor. Google uses BERT to enhance its search predictions. In our first study, we have taken the data from the dataset and run it with BERT model [35].

For BERT analysis the probability analysis can be represented using the following language model by Equation (1) [36]:

$$\mathrm{P}(\mathrm{w}_1, \mathrm{w}_2 \dots, \mathrm{w}_\mathrm{T}) = \prod_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{P}(\mathrm{w}_\mathrm{t} | \mathrm{w}_1, \mathrm{w}_2, \cdots \mathrm{w}_{\mathrm{t}-1}) \ [36] \qquad (1)$$

Where w_1, w_2 … are the different individual entities of which we need to find the probability distribution and T is the total number of entities. In our case, this is the probability distribution of each word in the subtitle file.

2) LDA: The Latent Dirichlet Allocation (or LDA) is a probabilistic model. The main aim of this model is to represent documents as different topics and each of these topics are characterised by a distribution over words [32]. The assumption made here is that every course has a set of topics already and the text (subtitles in our case) have relevant information to summarise these topics and hence, they can be grouped under them. The algorithm tells us the similarities in the data by grouping them into common topics [14]. It gives us a distribution of the word usage and when we search for a particular word, it refers to this distribution [33]. Supervised Machine Learning algorithms are used to run the model. This approach is used as a solution to a lot of problems related to topic identification, face recognition, web spam classification and entity resolution [34]. The second part of our study deals with LDA.

To find the normal probability density function using the LDA method, the formula is given by Equation (2) [37]:

$$P(X|\pi_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \, exp\left[-\frac{1}{2}(X-\mu_i)'\Sigma^{-1}\,(X-\mu_i)\right] \ [37](2)$$

Where,
πi – Probability density function
x – Multivariable normal
μi – Mean vector
Σ – Variance- covariance matrix.

This can be used if all the matrices for all the populations are homogenous. The decision rule of the LDA algorithm is based on the Linear score function, which is defined by Equation (3):

$$\mathrm{S}_\mathrm{i}^\mathrm{L}(\mathrm{x}) = -\frac{1}{2}\mu_\mathrm{i}'\Sigma^{-1}\mu_\mathrm{i} + \mu_\mathrm{i}'\Sigma^{-1}\mathrm{X} + \log \mathrm{P}(\pi_\mathrm{i}) \ \ [38] \qquad (3)$$

Where following substitutions are made:

$$d_{io} = -\frac{1}{2}\mu_i'\Sigma^{-1}\mu_i \qquad ; d_{ij} = jth\ element\ of\ \mu_i'\Sigma^{-1}$$

$\mathrm{d}_\mathrm{i}^\mathrm{L}(\mathrm{X})$ is the linear discriminant function (4) i.e.

$$\mathrm{d}_\mathrm{i}^\mathrm{L}(\mathrm{X}) = \mathrm{d}_\mathrm{io} + \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{d}_\mathrm{ij}\mathrm{x}_\mathrm{j} \qquad (4)$$

Therefore we get Equation (5) [38],

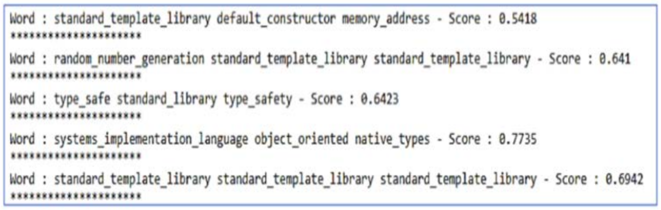
$$\mathrm{S}_\mathrm{i}^\mathrm{L}(\mathrm{x}) = \mathrm{d}_\mathrm{i}^\mathrm{L}(\mathrm{X}) + \log \mathrm{P}(\pi_\mathrm{i}) \ \ [38] \qquad (5)$$

IV. Experiment and Results

A. BERT

BERT analysis was first carried out on the preprocessed data. There was a restriction put on the number of concepts that could be extracted to only 3 concepts per line. The n-gram set for each concept was between 1 and 3. Fig. 1 depicts the results obtained.

Fig. 2 shows the coherence and the average overlap of the topics when the data was processed through the BERT model. 20 topics were given to derive BERT’s selected keywords. The topic coherence graph shows linear increase upto topics, which is also followed by a linear increase, but the slope gets reduced. The average topic overlap graph shows a steep linear decrease initially up to 2 topics, after which the slope reduces. Finally after 3 topics, the line almost flattens out. Both the graphs overlap at a point in the earlier stages. The ideal number of topics is 4.



```
Word : standard_template_library default_constructor memory_address - Score : 0.5418
*********************
Word : random_number_generation standard_template_library standard_template_library - Score : 0.641
*********************
Word : type_safe standard_library type_safety - Score : 0.6423
*********************
Word : systems_implementation_language object_oriented native_types - Score : 0.7735
*********************
Word : standard_template_library standard_template_library standard_template_library - Score : 0.6942
*********************
```

Fig. 1. The Concept Extracted from Subtitlles with BERT.

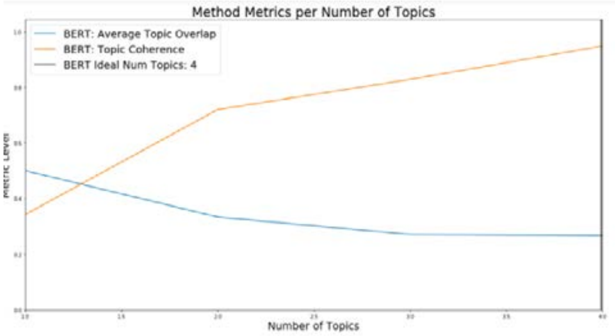
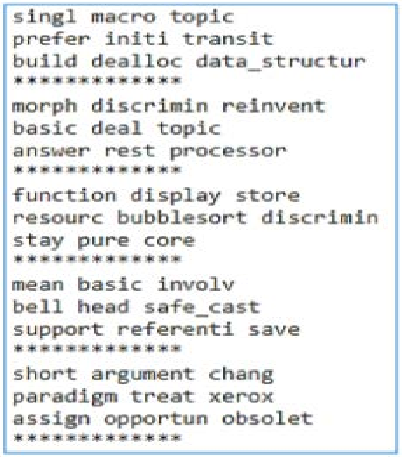


Fig. 2. BERT Coherence Score with Overlap Coefficient.

B. LDA

The second studies were carried out using the LDA model. Again, concepts per line were restricted to 3 and the n-gram was set between 1 and 3. Fig. 3 gives us the results obtained.



```
singl macro topic
prefer initi transit
build dealloc data_structur
*************
morph discrimin reinvent
basic deal topic
answer rest processor
*************
function display store
resourc bubblesort discrimin
stay pure core
*************
mean basic involv
bell head safe_cast
support referenti save
*************
short argument chang
paradigm treat xerox
assign opportun obsolet
*************
```

Fig. 3. The Concept Extracted from Subtitles with BERT.

Fig. 4 gives us an insight of the results for the same. A graph containing coherence, average overlap of topics was plotted where 20 topics were given to derive the LDA selected keywords. The graph of Topic Coherence shows a peculiar trend. It remains constant throughout and shows no variations at all. The average topic overlap shows a slight decreasing linear trend up to 2 topics, then it remains constant and above 3 topics there is a further linear slight decrease. The overall overlap decrease is very small. Unlike what we observed in the graph of BERT analysis, in this graph we do not see any intersection of the two parameters.

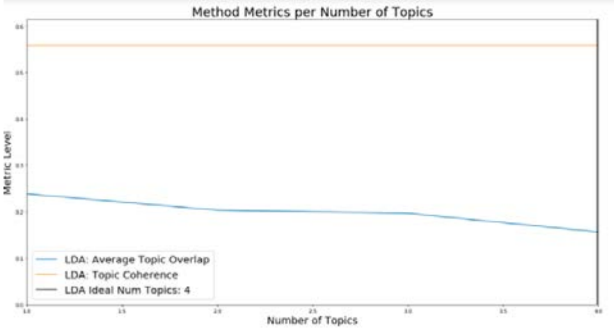


Fig. 4. LDA Coherence Score with Overlap Coefficient.

C. BERT v/s LDA

In order to draw comparisons with both the studies, the following parameters were considered.

1) Overlap coefficient: It is the measure of similarity that is used to track the amount of overlap between two finite sets. In other words, we can say that it is the intersection of two sets. Our studies showed the average overlaps of the topics in LDA to be higher than that of BERT analysis.

2) Topic coherence: It measures the total score of a single topic by measuring the degree of semantic similarity between the high scoring words of the topic. The consistency of the concepts by BERT was found to be higher than that of LDA.

As consistency is the more prioritised factor, overall, it can be concluded that BERT is the better model to use for keyphrase extraction of video subtitles in MOOC than LDA as it gives us clearer information about the topic coherence.

V. Conclusion and Future Prospects

Our studies show that BERT was a better model that could be implemented in order to extract keyphrases from the video subtitles from MOOC videos. The MOOC industry is booming and will continue to do so in the future. It is important to ensure that the course completion rate is high. Now that one can identify the key topics in a course using BERT model, further programming can be done to link these results with the search history of the learner. When any of the key topics are searched, these courses should show up and similar courses be recommended. This will ensure that the learner finds exactly what they are looking for thus motivating them to complete the course and enjoy it. This also helps give them personalised recommendations. As mentioned earlier, the teachers recording the courses also will vastly benefit from this. They can check the engagement of the students in their courses, or have an idea about which part of their video is watched more or gets more demand. They can also use this data to record lectures accordingly so that their courses appear on the top of the recommendations.

References

[1] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. Research & Practice in Assessment, 8, 13-25.

[2] Zhou, Y., & Xu, Z. (2020, August). Multi-Model Stacking Ensemble Learning for Dropout Prediction in MOOCs. In Journal of Physics: Conference Series (Vol. 1607, No. 1, p. 012004). IOP Publishing.

[3] Brinton, C. G., & Chiang, M. (2015, April). MOOC performance prediction via clickstream data and social learning networks. In 2015 IEEE conference on computer communications (INFOCOM) (pp. 2299-2307). IEEE.

[4] Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z., & Wong, F. M. F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model. IEEE transactions on Learning Technologies, 7(4), 346-359.

[5] Guo, P. J., Kim, J., & Rubin, R. (2014, March). How video production affects student engagement: An empirical study of MOOC videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 41-50).

[6] Kuh, G. D. (2003). What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. Change: The magazine of higher learning, 35(2), 24-32.

[7] Giannakos, M. N., Sampson, D. G., & Kidziński, Ł. (2016). Introduction to smart learning analytics: foundations and developments in video-based learning. Smart Learning Environments, 3(1), 1-9.

[8] Mohamad, N., Ahmad, N. B., & Sulaiman, S. (2017). Data pre-processing: a case study in predicting student's retention in MOOC. Journal of Fundamental and Applied Sciences, 9(4S), 598-613.

[9] Bach, S. H., Broecheler, M., Kok, S., & Getoor, L. (2010). Decision-driven models with probabilistic soft logic.

[10] [Wan, X., & Xiao, J. (2008, July). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In AAAI (Vol. 8, pp. 855-860).

[11] Gollapalli, S. D., & Caragea, C. (2014, June). Extracting keyphrases from research papers using citation networks. In Twenty-eighth AAAI conference on artificial intelligence.

[12] Wu, Y. F. B., Li, Q., Bot, R. S., & Chen, X. (2005, October). Domain-specific keyphrase extraction. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 283-284).

[13] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019, November). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 1441-1450).

[14] Krestel, R., Fankhauser, P., & Nejdl, W. (2009, October). Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems (pp. 61-68).

[15] Canini, K., Shi, L., & Griffiths, T. (2009, April). Online inference of topics with latent Dirichlet allocation. In Artificial Intelligence and Statistics (pp. 65-72). PMLR.

[16] Hasan, K. S., & Ng, V. (2014, June). Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1262-1273).

[17] Hasan, K. S., & Ng, V. (2010, August). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In Coling 2010: Posters (pp. 365-373).

[18] Albahr, A., Che, D., & Albahar, M. (2021). A novel cluster-based approach for keyphrase extraction from MOOC video lectures. Knowledge and Information Systems, 63(7), 1663-1686.

[19] Koka, R. S., Chowdhury, F. N., Rahman, M. R., Solorio, T., & Subhlok, J. (2020, December). Automatic identification of keywords in lecture video segments. In 2020 IEEE International Symposium on Multimedia (ISM) (pp. 162-165). IEEE.

[20] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

[21] Martinez - Romo, J., Araujo, L., & Duque Fernandez, A. (2016). S em G raph: Extracting keyphrases following a novel semantic graph - based approach. Journal of the Association for Information Science and Technology, 67(1), 71-82.

[22] Florescu, C., & Caragea, C. (2017, July). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1105-1115).

[23] Pan, L., Wang, X., Li, C., Li, J., & Tang, J. (2017, November). Course concept extraction in moocs via embedding-based graph propagation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 875-884).

[24] Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 239, 39-57.

[25] Terrizzano, I. G., Schwarz, P. M., Roth, M., & Colino, J. E. (2015, January). Data Wrangling: The Challenging Yourney from the Wild to the Lake. In CIDR.

[26] Mifrah, S., & Benlahmar, E. H. (2020). Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus. International Journal of Advanced Trends in Computer Science and Engineering, 5756-5761.

[27] Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval (Vol. 161175).

[28] Li, W. J., Wang, K., Stolfo, S. J., & Herzog, B. (2005, June). Fileprints: Identifying file types by n-gram analysis. In Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop (pp. 64-71). IEEE.

[29] Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?. In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics.

[30] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8, 842-866.

[31] Khodeir, N. A. (2021). Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT. IEEE Access, 9, 58243-58255.

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

[33] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[34] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.

[35] Bíró, I., Siklósi, D., Szabó, J., & Benczúr, A. A. (2009, April). Linked latent dirichlet allocation in web spam filtering. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 37-40).

[36] Dadure, P., Pakray, P., & Bandyopadhyay, S. (2021). BERT-Based Embedding Model for Formula Retrieval. CLEF.

[37] Jin, Q., & Waibel, A. (2000, October). Application of LDA to speaker recognition. In Interspeech (pp. 250-253).

[38] Mudde, R. F., Groen, J. S., & Van Den Akker, H. E. A. (1998). Application of LDA to bubbly flows. Nuclear Engineering and Design, 184(2-3), 329-338