Education and Information Technologies (2023) 28:5567–5584
https://doi.org/10.1007/s10639-022-11373-1

Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques

Hajar Zankadi[1] · Abdellah Idrissi[1] · Najima Daoudi[2] · Imane Hilal[2]

Received: 4 August 2022 / Accepted: 22 September 2022 / Published online: 4 November 2022

Abstract

Interests play an essential role in the process of learning, thereby enriching learners 'interests will yield to an enhanced experience in MOOCs. Learners interact freely and spontaneously on social media through different forms of user-generated content which contain hidden information that reveals their real interests and preferences. In this paper, we aim to identify and extract the topical interest from the text content shared by learners on social media to enrich their course preferences in MOOCs. We apply NLP pipeline and topic modeling techniques to the textual feature using three well-known topic models: Latent Dirichlet Allocation, Latent Semantic Analysis, and BERTopic. The results of our experimentation have shown that BERTopic performed better on the scrapped dataset.

Keywords Social Learning · MOOCs · Topic modelling · Course interest · Social media

✉ Hajar Zankadi
hajar_zankadi@um5.ac.ma

Abdellah Idrissi
idriab@gmail.com

Najima Daoudi
ndaoudi@esi.ac.ma

Imane Hilal
ihilal@esi.ac.ma

[1] Computer Science Department IPSS team, Faculty of sciences Rabat, Mohammed V University, Rabat, Morocco

[2] Lyrica Labs, Information Science School (ESI), Rabat, Morocco

1 Introduction

The implication of social media in MOOCs allows advocating an important place for collaboration and social interaction. Besides, social media has become a daily activity which leads to a growing reliance on it. It is characterized by its interactivity, connectedness, and user-generated content. Furthermore, social media plays a major role in sharing information through interactive communication that reflects users' thoughts, feelings, and behavior. Consequently, social media use and its impact on socialization constitute a viable source to accurately and effectively capture and analyze users' interests and preferences.

On the other hand, MOOCs provide an open network for exchanging and sharing learning materials at a low cost, regardless of time and location. Despite these key characteristics, MOOCs suffer from low completion rates and high drop-out rates. Additionally, the lack of interaction and satisfaction are considered major factors that are related to learners' decision to drop out (Zankadi et al., 2019).

In the context of MOOCs, Interests play an important role in the learning process as they are energizing it through a powerful motivational process that contributes to guiding the academic and career trajectories of the learners (McIntyre et al., 2021) (Harackiewicz et al., 2016). A learner's interest is a key component of adaptive hypermedia and educational systems that focus on the learner's behaviors and personalize courses according to learner interests (Peng et al., 2016). Learners are motivated to invest time and effort toward the course of their interest, thereby, enriching learners' interest will yield to a better discovery of course subjects that is the best fit with their preferences which impact their satisfaction and thereby their interaction inside MOOCs.

MOOCs already recommend courses that respond to learners' interests based on their participation. However, the same learners are more interactive in social media through the content they generate and which contains hidden information about their "real" interests and preferences. Since generating user interest is a challenging task, using topic modeling techniques is useful to uncover the main thematic information related to a user (Bai et al., 2021).

In our paper, we propose a Course Topic Model (CTM) based on Natural Language Processing (NLP) and topic modeling techniques to identify and detect learners' course of interest based on their spontaneous interaction in social media, in particular Twitter. The generated CTM contains the most probable topic of interest for each learner. We train three well-known models: Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and BERTopic after applying the NLP pipeline to the tweets shared by the learners. Then, we evaluate the models using topic coherence score, topic diversity score, and human judgment. Experiments were performed to reveal that BERTopic and LDA model performed better on the scrapped dataset and their results are used to generate the course topic model.

Following this introduction, we outline the concepts used and the problem statement of our work. In section III, we point out the related work. We then highlight the proposed approach in section IV. We discuss and evaluate the outcome of our approach in section V before presenting the conclusion and future work in section VI.

2 Context and problem statement

2.1 Concepts

- **Topic modeling**:

Topic modeling is an unsupervised technique for discovering the hidden topics in a collection of documents (Alami et al., 2021). It is used to analyze text documents and to extract automatically the latent themes from them (Sharma et al., 2021).

- **Social learning**:

Albert Bandura suggests that learning occurs in a social environment by observing and replicating what others do (Bandura & Walters, 1977). Social learning assumes iterative feedback between the learner and their environment (Castellanos-Reyes et al., 2021). It is the result of interactions between individuals with a common purpose (Anderson et al., 2020). Through various types of user-generated content (e.g., videos, images, audio, text posts), people actively connect and cooperate in the social media environment. As a result, users can learn how to act and behave by watching videos or images that other users in their network have created or shared (C.-T. Peng et al., 2019), and as users become learners, they can learn about underlying cultural conventions via text posts that articulate how other users solve problems.

2.2 Topic modeling techniques

The most popular topic modeling techniques are:

- **Latent Dirichlet Allocation (LDA)**:

LDA is an unsupervised clustering technique used for text analysis. it is first proposed by Blei et al. (Blei et al., 2003). as a generative probabilistic model in which each document is considered as a distribution of topics and each topic is considered as a distribution of words (Guo et al., 2021).

LDA model has two Dirichlet distributions: Alpha (ɑ) controls the document-topic distribution, and Beta (β) controls the topic-word distribution. Figure 1 represents the graphical model of LDA, where and as highlighted in (Gupta & Katarya, 2021):

- Θ_m is the topic mix for document m where Θm ~ Dirichlet(α).
- Zm,n is a topic assignment for word n in document m.
- Wm,n is word n in document m.
- N is the total number of words in all documents (the size of the vocabulary).
- M is the number of documents.
- K is the number of topics.
- Φ_k is the distribution of words for topic k where Φk ~ Dirichlet(β).

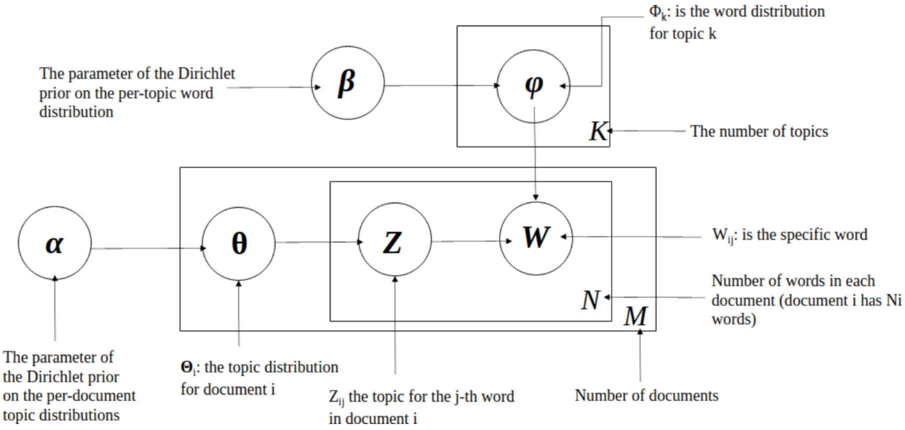
Φ_k: is the word distribution for topic k
The parameter of the Dirichlet prior on the per-topic word distribution
β
φ
K
The number of topics
α
θ
Z
W
W_{ij}: is the specific word
Number of words in each document (document i has Ni words)
N
M
The parameter of the Dirichlet prior on the per-document topic distributions
Θ_i: the topic distribution for document i
Z_{ij} the topic for the j-th word in document i
Number of documents

Fig. 1 Generative process of LDA

- **Latent Semantic Analysis (LSA)**:

Latent semantic analysis (LSA) is an unsupervised, indexing method in NLP to extract the words which are semantically related to each other (Qi et al., 2021). LSA begins by filling in a matrix with the word counts for each sentence or paragraph. Each row denotes a unique word, while each column denotes a sentence or paragraph (Suleman & Korkontzelos, 2021). The sentence-word correlation is found using SVD, a dimensionality reduction method, that is able to improve accuracy by reducing noise (Yang et al., 2022).

- **BERTopic**:

BERTopic is a topic modeling algorithm that generates topics as follow: first, converting each document to an embedding representation using BERT model. Then, lowering the dimensionality of the embeddings using UMAP to optimize the clustering process with HDBSCAN and finally, creating the topics using a custom class-based variation of TF-IDF (Grootendorst, 2022).

2.3 Problem statement

MOOCs have been growing rapidly thanks to their outstanding advantages of openness, unlimited access to resources, curriculum sharing, and low cost. They excelled across humanities, business management, health, and science-based subjects which elevate global enrolments and learners' participation via dedicated web portals, regardless of location.

MOOCs have gained a new level of attention during covid 19 pandemic and are considered a highly effective and flexible learning tool. They have proved to be a facilitator for the teaching-learning process and a replacement for the traditional face-to-face learning methods. Recently, around 2.8 K courses were added, and 16.3 K MOOCs were announced or launched by around 950 universities around the world

(Er-Rafyg et al., s. d.) which proves the high demand for MOOCs. However, two key challenges affect persistence in MOOCs: low course completion rate and high dropout rate. According to (Badali et al., 2022), over 90% of enrollees never finish the course and MOOCs' retention rate ranges between 3 and 15%. Many factors impact the learners' decisions in all stages of MOOCs acceptance, including feelings of isolation and lack of interaction, lack of motivation, learner characteristics, MOOCs features, and learner experience (Badali et al., 2022) (Gupta & Maurya, 2022).

Alternatively, interactivity and social presence are two important characteristics of social media that promote social connectedness, stickiness, information sharing, and collaboration among many users (Lin & Kishore, 2021). Social media has revolutionized how humans interact, providing them with opportunities to satisfy their social needs with respect to their interests and preferences (Kross et al., 2021).

In this context, providing learners with social learning that supports their interests and preferences will promote their engagement through increasing their role of sociability, sense of community, and course satisfaction (Yılmaz & Yılmaz, 2022). According to (Crane & Comley, 2021), social learning has been demonstrated to be effective to raise learners' satisfaction via reducing the feelings of isolation and lack of interactions. In (Liu et al., 2022), the authors make it clear that learner motivation—both intrinsic (interest) and extrinsic (perceived knowledge value—is essential to engagement and, consequently, to higher levels of learner satisfaction.

In our previous work (Zankadi et al., 2022), we implemented a social profile ontology that models the learner profile characteristics in both MOOCs and social media. We first created two local ontologies that describe the learner profile in MOOCs and social media, then we relied on ontology mapping and merging techniques to maintain the semantic links between the different concepts of the two implemented ontologies. The construction of the social profile is achieved by incorporating well-known standards and ontologies such as the IMS-LIP standard, FOAF, and SIOC ontology as well as the major concepts that describe the best user profile in social media. Interests and preferences are two important components in the social profile ontology that influence learner motivation and engagement.

Our main goal is to ensure persistence in MOOCs through social learning and by providing learners with courses that best match their interests and preferences. In this way, learners' satisfaction will increase as they will be more motivated to interact and participate inside MOOCs. In this work, we apply NLP and topic modeling techniques to the text content shared by learners on social media. The generated content contains hidden information that reveals the real interest and preferences of the learner.

3 Related work

Topic modeling techniques have been applied to several fields such as e-learning and social media.

3.1 Topic modeling in MOOCs

Topic modeling was used to detect learning topics of interest from course reviews in MOOCs. Authors (Liu et al., 2017) used an author topic model to extract the topic of interest of each learner based on the unstructured review data of MOOCs for personalized course recommendations. In another work (Peng et al., 2016), the authors proposed a Like-Latent Dirichlet Allocation (Like-LDA) model by incorporating the behavioral feature "like" to build the learner topic interest profile. Furthermore, the work presented in (Liu et al., 2019), pointed out a Behavior-Sentiment Topic Mixture (BSTM) topic model that incorporates emotion and behavior features to detect learners' oriented topics as well as learners' sentimental tendency and interaction toward these topics. Moreover, the authors (Lubis et al., 2019), developed a topic model using LDA, sentiment analysis and a sentence filtering approach based on helpful subjective reviews in MOOCs.

The topic modeling approach was applied as well on discussion forum posts in MOOCs. Authors in (Peng et al., 2020) analyzed the discussion posts of learners using a Time-Information-Emotion Behavior Model (TI-EBTM) to automatically track the temporal topic changes in the discussion forums. Similarly, the work highlighted in (B. Yang et al., 2022) used latent semantic analysis (LSA) to classify and identify learners with distinct longitudinal profiles of topic-relevant forum posting in MOOCs. Further, the Authors (Onan & Toçoğlu, 2021), proposed a document-clustering model using weighted word embedding and clustering approach to detect questions topics in MOOCs discussion forum posts.

3.2 Topic modeling in social media

The topic modeling approach was used in social media to detect user interest tags. In (He et al., 2020), authors applied a Bi-Labeled LDA model to automatically extract interest tags for non-famous users inferred from the topical interest of famous users in social media. Further, authors in (Chen & Ren, 2017) presented an extension of LDA model called Forum-LDA model to capture coherent topics and serious interests of the users in a forum for the aim of recommendation. Moreover, authors (Kim et al., 2021), proposed a framework based on LDA model that incorporates both textual and visual features to discover user interests as well to recommend points-of-interests (POIs) and potential friends to the target user. In addition, the work in (Yu & Li, 2021) highlighted a method to identify the interest themes of users in microblog using a multi-granular text feature representation including text vector, an LDA model to extract the topic tags, and LSTM to learn the semantic features of the sentences.

Topic modeling was applied for sentiment analysis as well. The work pointed out in (Heidari & Jones, 2020), introduced a new BERT model for sentiment classification of tweets to independently identify topic features for the social media bot. A deep learning-based topic-level sentiment analysis model was proposed by (Pathak et al., 2021) and applied on streaming short text data in social media.

To synthesize, topic modeling has been widely used to predict user interest from the textual feature in both MOOCs and social media with different purposes and

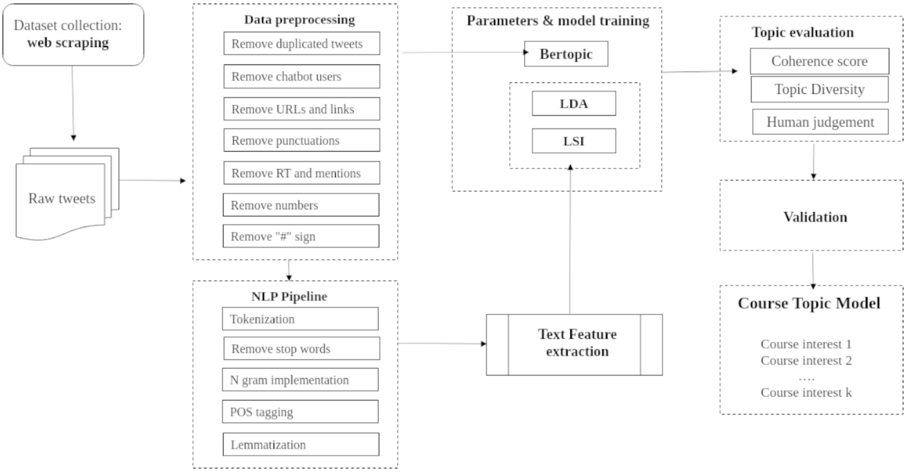

Fig. 2 Sequence of steps for the proposed approach

objectives. To our knowledge, there have not been any works in literature that apply the topic modeling approach to learners' interactive content in social media to detect and extract their course of interest.

We aim to explore the learner's spontaneous interaction and the generated content in social media to reveal his real interests and preferences and enrich them with the already existing interests in MOOCs by applying NLP and topic modeling techniques to textual features shared by the learner.

The following sections describe in more detail our approach and experimentation.

4 Our approach

Our goal is to apply NLP and Topic modeling techniques to the tweets generated by the learners on social media to identify their "real" course of interest.

The proposed approach encompasses several steps related to the generation of topical interests as highlighted in Fig. 2. The steps are performed sequentially for data preparation, data preprocessing, and feature extraction. Then we train three different topic modeling methods, including LDA, LSA, and BERTopic, and finally, we evaluate the performance of the models using topic evaluation metrics and validate their coherence.

The sequence of steps in the proposed approach is discussed in detail below:

4.1 Dataset collection

A set of tweets were collected via web scraping technique for two months. We used Twitter API keys and Netlytic[1] which is a cloud-based text and social networks analyzer, that allows us to collect publicly accessible posts from social media.

[1] https://netlytic.org/index.php.

To refine the process of scraping, we applied two filters: scarping data in the English language and initiating the scarping using the keywords «Computer science» and «Artificial intelligence ». A total of 2500 tweets per query were scraped. The dataset used for our experimentation contained 120 000 tweets from 12,187 learners. The meta-data gathered included tweet id, learner names, tweet text, tweet type, learner id, etc. Since we are interested in the textual feature generated by the learners, we kept only learners' names and their tweets as features for our dataset.

4.2 Data preprocessing

We first cleaned our dataset from duplicated tweets and chatbot users using *string pattern recognition*. Each tweet should be free of links, numbers, emojis, and punctuation. We used a preprocessing library called ***tweet-preprocessor*** to convert our tweets into the processed form.

Preprocessing steps included: *removing URLs, removing Mentions, removing Emojis and Smileys, and removing Numbers*. For hashtags, we removed only the sign "#" and we kept the word after the sign since it constitutes valuable information about the preferences of the learner. We removed extra white spaces and punctuation as well.

4.3 NLP pipeline

The following steps were performed:

- ***Tokenization***: we break the sentences into words, lowercasing the words, ignoring tokens that are too short (contains at least 3 letters), and removing letter accents.
- ***Remove stop words***: Remove all commonly used words like "a", "an", "the", etc. As they do not contribute to either the representation of the tweet or to the scoring mechanism of each tweet. We also extend the list of the stop words by high-frequency words. This is done by setting a threshold in which words that occur more frequently than the threshold should be removed. A standard stop word list by the "NLTK" library (Siva Rama Rao et al., 2022) has been used for this work.
- ***N-grams implementation***: we extract a sequence of 'n-items' that occur frequently in a sentence. In our work, we implement bi-grams (2 words), and tri-grams (3 words). The Phraser model of the Gensim library (Haider et al., 2020) has been used to construct the bi-gram and tri-gram models.
- ***POS / Speech of tag selection***: also referred to as the processing pipeline. The default pipeline consists of a tagger, a parser, and an entity recognizer. In our work, we only use a tagger. We check part of the speech tag of each token and keep the nouns, adjectives, verbs, and adverbs.
- ***Lemmatization***: We return the base or dictionary form of a word, known as the lemma, and merely remove inflectional endings.

We limited our data to a sample of 10,000 tweets since training the models with 120 000 tweets - was time and memory-consuming.

The output of this step is a clean and tokenized set of terms that will be used as an input for the following steps.

4.4 Text feature extraction

Text feature extraction is the mechanism of transforming text documents into vector representations. In our work, each tweet represents one document. We have 10 000 documents in total. We applied feature extraction using Bag of words and TF-IDF transformations to generate our corpora.

A Bag of Words is a representative model of document data, that highlights the occurrence of words within a document (Diera et al., 2022) while a TF-IDF transformation provides a weight for each word. The value of TF-IDF is determined by multiplying two metrics: the relative frequency of a word in a specific document and the inverse proportion of the word across the entire set of documents, which reflects how relevant a word is to a particular document (Zaware et al., 2021).

To train both LDA and LSI, a dictionary and a corpus should be created. A dictionary encapsulates the mapping between normalized words and their integer ids. On the other hand, the corpus is a list of lists containing tuples for each word id and its frequency.

While creating the dictionary, we apply the following steps:

- Filtering out high-frequency words: we used the ***FreqDist*** function to create the frequency distribution of all the words in the text. We set a threshold equal to 2000 to remove words that have frequency over them.
- Adding those words to the stop words list.
- Performing the ***filter_extremes***[2] function that filters out tokens in the dictionary by their frequency. It has two variables: "***no_below***" and "***no_above***". This means to keep tokens contained in at least no_below documents and no more than no_above documents. We repeated this process using respectively a set of test values:({5,0.5}, {5,0.7}, {5,1}), ({10, 0.5}, {10,0.7}, {10,1}) and ({20, 0.5}, {20,0.7}, {20,1}). The best result was obtained using no_below=5 and no_above=0.7.
- Based on the result above, we remove all words that occur in less than 5 documents (tweets) and all words that occur in more than 70% of all the documents.

Based on the created dictionary, two corpora were generated: BOW corpus and TFIDF corpus.

4.5 Parameters and model training

We apply LDA and LSI to both corpora. There are two types of hyperparameters — the number of latent topics for both models and Dirichlet priors (alpha and beta) for LDA.

[2] https://tedboy.github.io/nlps/generated/generated/gensim.corpora.Dictionary.filter_extremes.html.

Table 1 LDA and LSA Models' hyperparameters

	num_topics	chunksize	passes	iterations	alpha	eta
LDA	10	2000	20	400	auto	auto
LSI	10	2000	-	-	-	-

Table 2 BERTopic model' hyperparameters

low_memory	calculate_probabilities	verbose	n_gram_ range	nr_ topics
True	True	True	(1,3)	auto

The number of latent topics must be chosen before LDA and LSI are run. We randomly choose a value of 10 for the number of latent topics. After, we perform a grid search over this hyperparameter to find the optimal number of topics as pointed out in the next section.

Alpha and beta are the Dirichlet priors for the LDA model. We set their values to **"Auto"** and the model learns the right values of these parameters when it is run.

The other parameters of LDA and LSI were chosen based on a set of test values that made the models converge. Table 1 highlights the values of the different hyperparameters.

For BERTopic, we use a Countvectorizer transformer to eliminate English stop words and ignore terms that have a document frequency strictly higher than 60 and lower than 20 to reduce the size of the resulting sparse c-TF-IDF matrix.

As a sentence-transformer model, we use BERTweet, a language model pre-trained for English Tweets. Table 2 points out the values chosen for the different parameters of BERTopic.

To prevent blowing up the memory in UMAP, we set "low_memory" to "True". The parameter "calculate_probabilities" is set to True as well to calculate the probability of a document belonging to any topic. In addition, we set the parameter "verbose" to True to track the stages of the model. For the number of topics, we set its value to "**auto**".

In the following section, we will describe the topic evaluation metrics used to validate the results of the trained models.

4.6 Topics evaluation

We evaluate and compare the performance of the different topic models. We use three evaluation metrics: coherence score, Inverted Rank-Biased Overlap (RBO) score, and human judgment.

The coherence score is defined as the median of pairwise word similarities formed by the top words of a given topic (Rosner et al., 2014). It measures the degree of semantic similarity between its high-scoring words. While Inversed Rank-Biased Overlap score evaluates how diverse the topics generated by a single model are. It compares the top N words of two topics and uses weighted ranking (Bianchi et al., 2020). A diversity close to 0 represents a redundant topic, and those close to 1

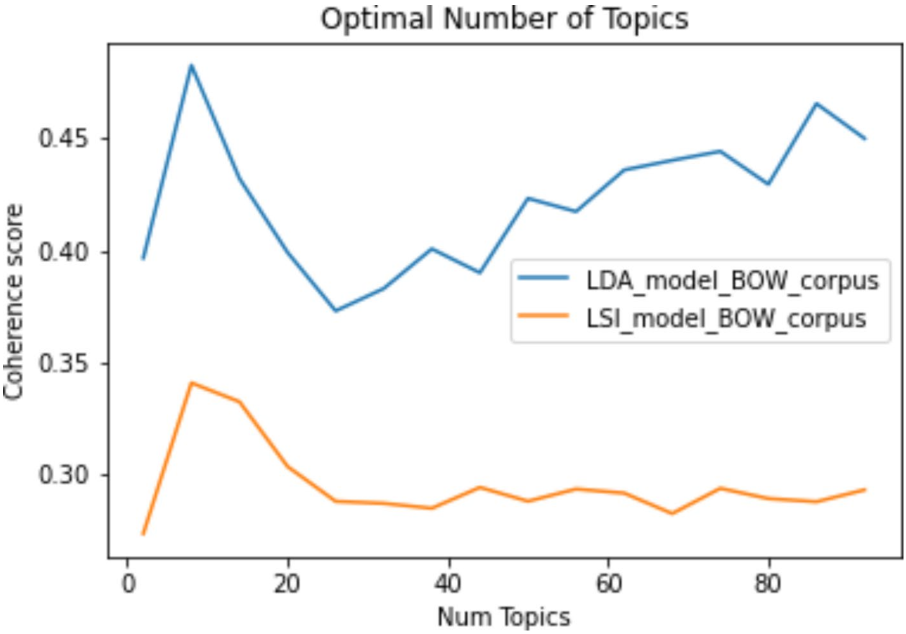

Fig. 3 Optimal number of topics for LDA and LSA models using BOW corpus

indicate more varied topics. The higher these metrics are, the better (Murakami & Chakraborty, 2022).

The approach adopted to find the optimal number of topics for LDA and LSI models is to build many models (LDA and LSI) with different values for a number of topics (k) and pick the one with the highest coherence value concerning the significance of the generated topics.

The search for the optimal number of topics started with a range from two to 98, with a step of 6. During the process, only one hyperparameter varied (number of topics) and the other remained unchanged until reaching the highest coherence score. Figures 3 and 4 present the values of the coherence score about the number of topics for LDA and LSA models with BOW and TFIDF corpora.

The optimal number of topics of each model is the one that has the highest coherence score concerning the topic's relevance and significance. The training of the BERTopic model results in several topics equal to **40** with a coherence score equal to **0,62.** Table 3 presents the optimal number of topics for each model and the corresponding coherence score.

Based on the above table, LDA and BERTopic models have the highest coherence score compared to LSA. To decide which model to adopt for the generation of the course topics, we compute the RBO score of LDA and BERTopic models as highlighted in Table 4 using OCTIS, a framework for training, analyzing, and comparing Topic Models (Terragni et al., 2021).

To judge the relevance of the topics and facilitate a clear interpretation of the extracted information from fitted LDA and BERTopic models, word cloud represen-

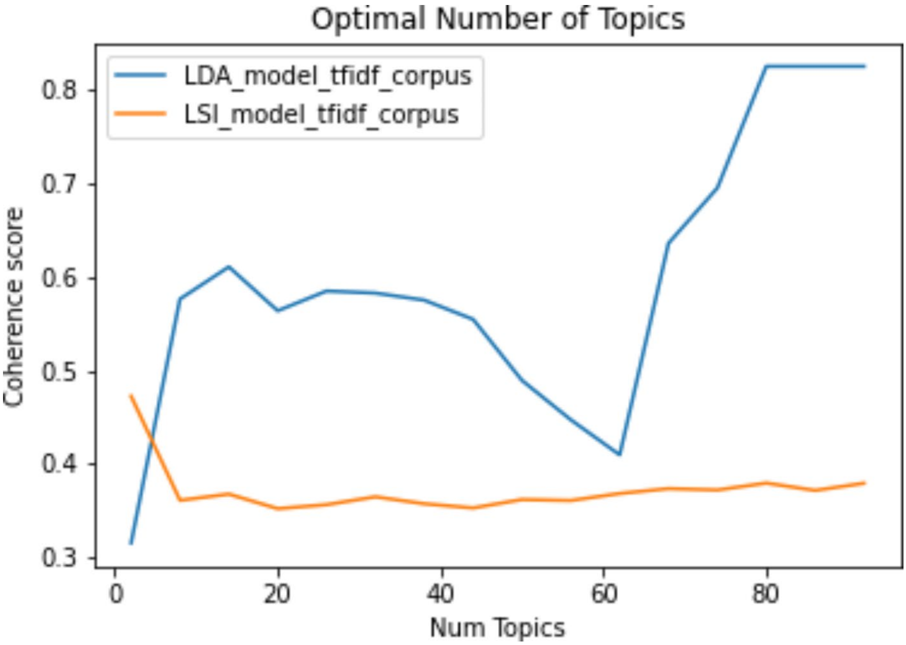

Fig. 4 Optimal number of topics for LDA and LSA models using TFIDF corpus

Table 3 Optimal number of topics with the corresponding coherence score

	BOW_LDA	BOW_LSA	TFIDF_LDA	TFIDF_LSA	BERTopic
Coherence score	0.50	0.34	0.59	0.48	0,61
The optimal number of topics	8	8	14	2	40

Table 4 RBO score of LDA and BERTopic models

	LDA_BOW	LDA_TFIDF	BERTopic
RBO	1	1	0.85

tation was used to generate a screenshot of the topics as pointed out in Figs. 5, 6 and 7, and 8.

We manually labeled the first five topical interests as highlighted in Tables 5 and 6, and 7 based on the word cloud representation for each model.

5 Discussion and results

The aim of this work is to predict the real interest of learners from the textual content they share on their social media account to enrich their course preferences in MOOCs. Although MOOCs offer a set of courses that suit the learners predefined interest when they first sign up but those preferences do not reflect their real needs and requirements.

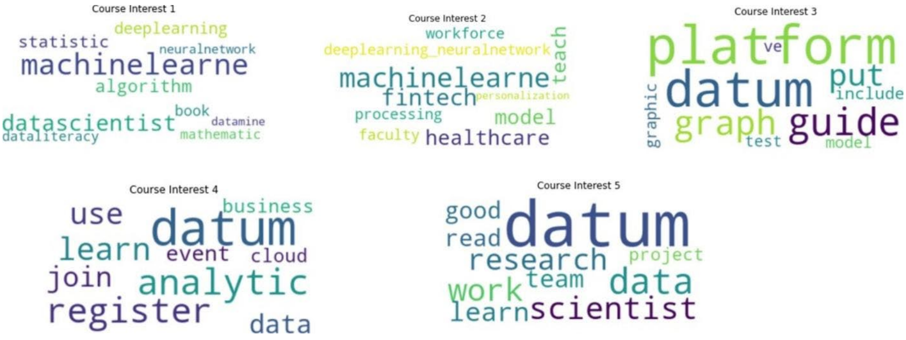

Fig. 5 Word Cloud representation of LDA topics with BOW corpus

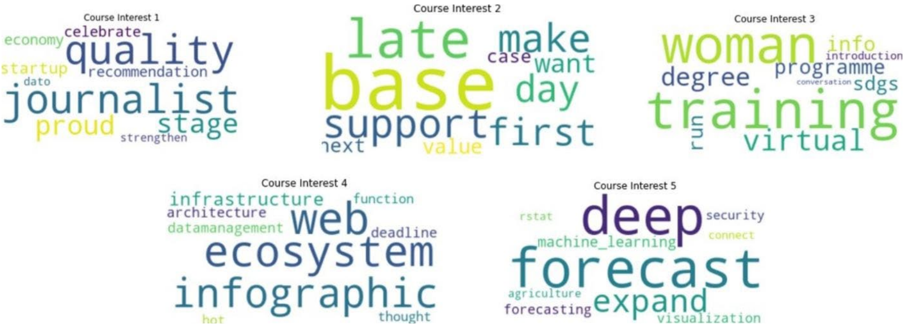

Fig. 6 Word Cloud representation of LDA topics with TFIDF corpus

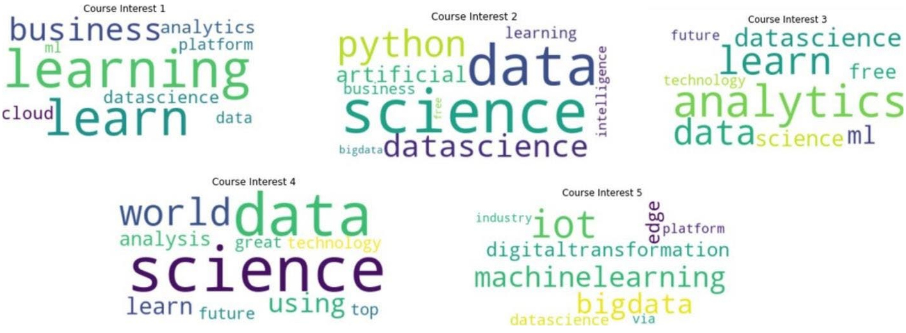

Fig. 7 Word Cloud representation of BERTopic topics

Table 5 Labels of the course interest generated from the LDA _BOW model

	Course Interest 1	Course Interest 2	Course Interest 3	Course Interest 4	Course Interest 5
LDA_BOW	Machine learning	Machine learning in fintech	Graph data platform	Data analytic	Datum research

Table 6 Labels of the course interest generated from the LDA _TFIDF model

	Course Interest 1	Course Interest 2	Course Interest 3	Course Interest 4	Course Interest 5
LDA_TFIDF	Journalism	--	Training Program for women	Digital ecosystem	Forecasting with machine learning

Table 7 Labels of the course interest generated from the BERTopic model

	Course Interest 1	Course Interest 2	Course Interest 3	Course Interest 4	Course Interest 5
BERTopic	Digital Business	Data Science with Python	Data Analytics	Data science and analysis	Industrial IoT

Table 8 Top 2 courses of interest of four learners

Learner ID	Top 2 courses of interest
ID 1	Digital business and data analytics
ID 2	Data science and machine learning
ID 3	Industrial IoT and digital business
ID 4	Artificial intelligence and data science

Our goal is to identify the course topics based on learners' interaction in Twitter, through the tweets they share. In order to do this, we require three models: LDA, LSI and BERTopic. We trained the LDA and LSI using two corpora: a bow corpus and a TFIDF corpus. We trained BERTopic using a pre-trained sentence transformer called BERTweet. We evaluated our models using the coherence score, the RBO score as well as human judgement to outline the quality and the relevance of the generated course topics.

LDA_BOW, LDA_TFIDF, and BERTopic models show prominent results with a coherence score of **0.50**, **0.59**, and **0.61** respectively, an RBO score of **1**, **1**, and **0 0.86**, respectively. To decide the best model to adopt for the course topics, we judge the relevance and the significance of the generated topics as well. We manually labeled the first five topical interests as highlighted in Tables 6, 7 and 8.

Based on this tables, BERTopic outperformed LDA regarding the quality and the relevance of the topical interest that are more representative for the course of interest of learners.

For each tweet, we identify the dominant topic that has the highest probability, and then for each learner, we get the top two dominant topics. Table 8 points out the three courses of interest for four learners.

6 Conclusion

Interaction is a key enabler of social media which allows users to spontaneously and freely socialize and collaborate through multiple forms of user-created content (e.g., videos, images, audio, text posts). This allows generating content that hides information about the real interests and preferences of the user. In contrast, MOOCs still suffer from low completion rates and high dropout rates that are related significantly to the lack of interaction and learner satisfaction.

Learners are motivated and satisfied to invest time and effort toward courses of their preferences, thereby, identifying learners' interests will encourage them to be committed to finishing their courses.

Our objective is to explore the aspect of stickiness and sociability of social media through the content generated to predict the real interest and preferences of the users using NLP and topic modeling techniques.

Our approach predicts automatically the real course of interest of the learners to enrich them with their already existing interests in MOOCs. We use for this purpose three well-known models: LDA and LSA with BOW and TFIDF corpora and BERTopic with BERTweet as sentence embeddings. Before training the models, we apply the NLP pipeline for data cleaning and preprocessing. We evaluate the models using the coherence score and RBO score and we judge the relevance of the topical interests by manually labeling them. Results have demonstrated that BERTopic outperformed the other models.

Our perspective is to develop an API that extracts and labels automatically the course of interest from the textual content in social media. The interesting tag generated automatically will serve to enrich the "interest component" of our social profile ontology that we had implemented in previous work. We aim as well to build a recommender system that takes into consideration the identified preferences of the learners to either recommend them courses when they first sign up in MOOCs using their social media account or to better orient them toward courses that responds to their profile's needs. Additionally, we aspire to evaluate our approach in a reel MOOC to major the reel impact of stickiness in MOOCs.

Declarations

Conflict of Interest None.

References

Alami, N., Meknassi, M., En-nahnahi, N., El Adlouni, Y., & Ammor, O. (2021). Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling. *Expert Systems with Applications*, *172*, 114652. https://doi.org/10.1016/j.eswa.2021.114652

Anderson, V., Gifford, J., & Wildman, J. (2020). An evaluation of social learning and learner outcomes in a massive open online course (MOOC): A healthcare sector case study. *Human Resource Development International*, *23*(3), 208–237. https://doi.org/10.1080/13678868.2020.1721982

Badali, M., Hatami, J., Banihashem, S. K., Rahimi, E., Noroozi, O., & Eslami, Z. (2022). The role of motivation in MOOCs' retention rates: A systematic literature review. *Research and Practice in Technology Enhanced Learning*, *17*(1), 1–20. https://doi.org/10.1186/s41039-022-00181-3

Bai, X., Zhang, X., Li, K. X., Zhou, Y., & Yuen, K. F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, *102*, 11–24. https://doi.org/10.1016/j.tranpol.2020.12.013

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (1 vol.). Prentice-hall Englewood Cliffs, NJ

Bianchi, F., Terragni, S., & Hovy, D. (2020). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*. https://doi.org/10.48550/arXiv.2004.03974

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022

Castellanos-Reyes, D., Maeda, Y., & Richardson, J. C. (2021). THE RELATIONSHIP BETWEEN SOCIAL NETWORK SITES AND PERCEIVED LEARNING AND SATISFACTION FOR EDUCATIONAL PURPOSES.Social Media: Influences on Education,231

Chen, C., & Ren, J. (2017). Forum latent Dirichlet allocation for user interest discovery. *Knowledge-Based Systems*, *126*, 1–7. https://doi.org/10.1016/j.knosys.2017.04.006

Crane, R. A., & Comley, S. (2021). Influence of social learning on the completion rate of massive online open courses. *Education and Information Technologies*, *26*(2), 2285–2293. https://doi.org/10.1007/s10639-020-10362-6

Diera, A., Lin, B. X., Khera, B., Meuser, T., Singhal, T., Galke, L., & Scherp, A. (2022). Bag-of-Words vs. Sequence vs. Graph vs. Hierarchy for Single-and Multi-Label Text Classification. *arXiv preprint arXiv:2204.03954*. https://doi.org/10.48550/arXiv.2204.03954

Er-Rafyg, A., Abourezq, M., Idrissi, A., & Bouhouch, A. (s. d.). *Courses Recommendations using Skyline BNL Algorithm*. 19. *International Journal of Artificial Intelligence.*

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*. https://doi.org/10.48550/arXiv.2203.05794

Guo, C., Lu, M., & Wei, W. (2021). An improved LDA topic modeling method based on partition for medium and long texts. *Annals of Data Science*, *8*(2), 331–344. https://doi.org/10.1007/s40745-019-00218-3

Gupta, A., & Katarya, R. (2021). PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. *Computers in biology and medicine*, *138*, 104920. https://doi.org/10.1016/j.compbiomed.2021.104920

Gupta, K. P., & Maurya, H. (2022). Adoption, completion and continuance of MOOCs: A longitudinal study of students' behavioural intentions. *Behaviour & Information Technology*, *41*(3), 611–628. https://doi.org/10.1080/0144929X.2020.1829054

Haider, M. M., Hossin, M. A., Mahi, H. R., & Arif, H. (2020). Automatic text summarization using gensim word2vec and k-means clustering algorithm. *IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 283–286, doi: https://doi.org/10.1109/TENSYMP50017.2020.9230670

Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy insights from the behavioral and brain sciences*, *3*(2), 220–227. https://doi.org/10.1177%2F2372732216655542

He, J., Liu, H., Zheng, Y., Tang, S., He, W., & Du, X. (2020). Bi-labeled LDA: Inferring interest tags for non-famous users in social network. *Data Science and Engineering*, *5*(1), 27–47. https://doi.org/10.1007/s41019-019-00113-0

Heidari, M., & Jones, J. H. (2020). Using bert to extract topic-independent sentiment features for social media bot detection. *11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2020, pp. 0542–0547, doi: https://doi.org/10.1109/UEMCON51285.2020.9298158

Kim, K., Kim, J., Kim, M., & Sohn, M. (2021). User interest-based recommender system for image-sharing social media. *World Wide Web*, *24*(3), 1003–1025. https://doi.org/10.1007/s11280-020-00832-9

Kross, E., Verduyn, P., Sheppes, G., Costello, C. K., Jonides, J., & Ybarra, O. (2021). Social media and well-being: Pitfalls, progress, and next steps. *Trends in Cognitive Sciences*, *25*(1), 55–66. https://doi.org/10.1016/j.tics.2020.10.005

Lin, X., & Kishore, R. (2021). Social media-enabled healthcare: A conceptual model of social media affordances, online social support, and health behaviors and outcomes. *Technological Forecasting and Social Change*, *166*, 120574. https://doi.org/10.1016/j.techfore.2021.120574

Liu, S., Ni, C., Liu, Z., Peng, X., & Cheng, H. N. (2017). Mining individual learning topics in course reviews based on author topic model. *International Journal of Distance Education Technologies (IJDET)*, *15*(3), 1–14. DOI: https://doi.org/10.4018/IJDET.2017070101

Liu, S., Peng, X., Cheng, H. N., Liu, Z., Sun, J., & Yang, C. (2019). Unfolding sentimental and behavioral tendencies of learners' concerned topics from course reviews in a MOOC. *Journal of Educational Computing Research*, *57*(3), 670–696. https://doi.org/10.1177%2F0735633118757181

Liu, Y., Zhang, M., Qi, D., & Zhang, Y. (2022). Understanding the role of learner engagement in determining MOOCs satisfaction: A self-determination theory perspective. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2022.2028853

Lubis, F. F., Rosmansyah, Y., & Supangkat, S. H. (2019). Topic discovery of online course reviews using LDA with leveraging reviews helpfulness. *International Journal of Electrical and Computer Engineering*, *9*(1), 426. DOI: https://doi.org/10.11591/ijece.v9i1.pp426-438

McIntyre, M. M., Gundlach, J. L., & Graziano, W. G. (2021). Liking guides learning: The role of interest in memory for STEM topics. *Learning and Individual Differences*, *85*, 101960. https://doi.org/10.1016/j.lindif.2020.101960

Murakami, R., & Chakraborty, B. (2022). Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts. *Sensors (Basel, Switzerland)*, *22*(3), 852. https://doi.org/10.3390/s22030852

Onan, A., & Toçoğlu, M. A. (2021). Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. *Computer Applications in Engineering Education*, *29*(4), 675–689. https://doi.org/10.1002/cae.22252

Pathak, A. R., Pandey, M., & Rautaray, S. (2021). Topic-level sentiment analysis of social media data using deep learning. *Applied Soft Computing*, *108*, 107440. https://doi.org/10.1016/j.asoc.2021.107440

Peng, C. T., Wu, T. Y., Chen, Y., & Atkin, D. J. (2019). Comparing and modeling via social media: The social influences of fitspiration on male instagram users' work out intention. *Computers in Human Behavior*, *99*, 156–167. https://doi.org/10.1016/j.chb.2019.05.011

Peng, X., Han, C., Ouyang, F., & Liu, Z. (2020). Topic tracking model for analyzing student-generated posts in SPOC discussion forums. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–22. https://doi.org/10.1186/s41239-020-00211-4

Peng, X., Liu, S., Liu, Z., Gan, W., & Sun, J. (2016). Mining learners' topic interests in course reviews based on like-LDA model. *International Journal of Innovative Computing Information and Control*, *12*(6), 2099–2110

Qi, Q., Hessen, D. J., & van der Heijden, P. G. (2021). A Comparison of Latent Semantic Analysis and Correspondence Analysis for Text Mining. *arXiv preprint arXiv:2108.06197*. https://doi.org/10.48550/arXiv.2108.06197

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*. https://doi.org/10.48550/arXiv.1403.6397

Sharma, A., Rana, N. P., & Nunkoo, R. (2021). Fifty years of information management research: A conceptual structure analysis using structural topic modeling. *International Journal of Information Management*, *58*, 102316. https://doi.org/10.1016/j.ijinfomgt.2021.102316

Siva Rama Rao, A. V., Vamsi, P., Rashmika, N., Hemanth, K., & Kumar, A. (2022). K. *Named Entity Recognition Using Stanford Classes and NLTK*. 583–597.doi: https://doi.org/10.1007/978-981-16-7657-4_47

Suleman, R. M., & Korkontzelos, I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications*, *165*, 114130. https://doi.org/10.1016/j.eswa.2020.114130

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). *OCTIS: Comparing and Optimizing Topic models is Simple!* 263–270. doi: https://doi.org/10.18653/v1/2021.eacl-demos.31

Yang, B., Tang, H., Hao, L., & Rose, J. R. (2022). Untangling chaos in discussion forums: A temporal analysis of topic-relevant forum posts in MOOCs. *Computers & Education*, *178*, 104402. https://doi.org/10.1016/j.compedu.2021.104402

Yang, X., Yang, K., Cui, T., Chen, M., & He, L. (2022). A Study of Text Vectorization Method Combining Topic Model and Transfer Learning. *Processes*, *10*(2), 350. https://doi.org/10.3390/pr10020350

Yılmaz, F. G. K., & Yılmaz, R. (2022). Exploring the role of sociability, sense of community and course satisfaction on students' engagement in flipped classroom supported by facebook groups. *Journal of Computers in Education*, 1–28. https://doi.org/10.1007/s40692-022-00226-y

Yu, Y., & Li, B. (2021). Microblog User Interest Recognition Based on Multi-Granularity Text Feature Representation. *In The 2nd International Conference on Computing and Data Science (pp. 1–10).*https://doi.org/10.1145/3448734.3450886

Zankadi, H., Hilal, I., Daoudi, N., & Idrissi, A. (2019). Towards a social learning environment. *In Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services* (pp. 607–610). https://doi.org/10.1145/3366030.3366120

Zankadi, H., Hilal, I., Idrissi, A., & Daoudi, N. (2022). A Social Profile Ontology to Enhance Learner Experience in MOOCs. *International Journal of Emerging Technologies in Learning*, *17*(4), https://doi.org/10.3991/ijet.v17i04.27389

Zaware, S., Patadiya, D., Gaikwad, A., Gulhane, S., & Thakare, A. (2021). Text summarization using tf-idf and textrank algorithm. *5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1399–1407, doi: https://doi.org/10.1109/ICOEI51242.2021.9453071

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.