Innovations in Education and Teaching International

ISSN: 1470-3297 (Print) 1470-3300 (Online) Journal homepage: www.tandfonline.com/journals/riie20

The untapped promise of educational data on social media: A machine learning approach to unveiling online learners

Yang Liu, Jingjing Zhang, Yehong Yang & Xudong Tao

To cite this article: Yang Liu, Jingjing Zhang, Yehong Yang & Xudong Tao (15 Sep 2025): The untapped promise of educational data on social media: A machine learning approach to unveiling online learners, Innovations in Education and Teaching International, DOI: 10.1080/14703297.2025.2561038

To link to this article: https://doi.org/10.1080/14703297.2025.2561038

Published online: 15 Sep 2025.

INNOVATIONS IN EDUCATION AND TEACHING INTERNATIONAL
https://doi.org/10.1080/14703297.2025.2561038

Routledge
Taylor & Francis Group

The untapped promise of educational data on social media: A machine learning approach to unveiling online learners

Yang Liu[a], Jingjing Zhang[b], Yehong Yang[b] and Xudong Tao[b]

[a]Institute of Education, China University of Geosciences, Beijing, China; [b]Faculty of Education, Beijing Normal University, Beijing, China

ABSTRACT
Massive Open Online Courses (MOOCs) attract millions of learners who engage in discussions about online learning on social media, creating a rich source of educational big data. However, current educational data mining often overlooks the broader learner demographics of learners across multiple courses. This study bridges this gap by analysing Weibo posts from 2013 to 2022 related to the MOOC phenomenon. Using machine learning and deep learning algorithms, learners were identified, with TextCNN proving the most effective. The Latent Dirichlet Allocation (LDA) topic model categorised post content, and Dynamic Topic Models (DTM) tracked the evolution of these topics over time. This research provides a novel perspective on online learners through social media, offering insights into diverse learning communities, identifying often-overlooked silent learners, and presenting a comprehensive view of the online learning experience. The findings lay a crucial foundation for integrating social media data into educational big data to improve online teaching and learning outcomes.

KEYWORDS
Moocs learner; social media; silent learners; text classification; topic modeling

Introduction

The public's reaction to the rise of Massive Open Online Courses (MOOCs) as a new educational phenomenon has been unprecedented, extending far beyond conventional news channels. Much of this discussion and debate about MOOCs has taken place on social media. While the most noticeable consequence of MOOCs so far has been to spark a multitude of new discussions on the state of online learning in the digital era, these conversations represent a valuable untapped data source.

Social media platforms have increasingly become a source of data for educational big data research. Among various educational studies utilising social media data, some have explored the use of X (formerly known as Twitter) as a primary means of interacting with students (Ansari & Khan, 2020; M. Liu et al., 2016) and delivering MOOC content (Mujahid et al., 2021). Other research has examined different aspects of MOOCs, such as the learner experience (Theophilou et al., 2024), and reflections on courses shared on Twitter (Khalil & Belokrys, 2022). Learners are frequently unwilling to report their experiences in online surveys or interviews; such datasets may therefore lack specific information regarding

CONTACT Jingjing Zhang jingjing.zhang@bnu.edu.cn School of Educational Technology, Beijing Normal University, Room 1302, Liyun Building 8, Hai-Dian District, Beijing, China

learners' perceptions and motivations. With a few significant exceptions – such as research on motivation and learning strategy (Barthakur et al., 2021; Fan et al., 2022; Wei et al., 2023) and on self-regulated learning – it is difficult to infer learners' authentic experiences with MOOCs from the limited data collected within the formal MOOC platforms. Typically, only students who actively participate in the course activities are represented, including in under-represented surveys and interviews. Due to the limited data available from the course platform, alternative sources, including social media, should be used to triangulate results by discovering and including silent learners in analyses.

Studies that compile data from multiple MOOCs are now rising in the area of big data educational research. Some studies have explored aspects like the timeliness of MOOCs (Zhang et al., 2015), learners' sentiments towards courses (Shen & Kuo, 2015), and discussions on the MOOC phenomenon (Costello et al., 2017). Despite these promising initiatives that use social media data as an additional dataset to address topics emerging on course platforms, there is still a dearth of research that makes effective use of existing large-scale social media collections. This may be due to a lack of meaningful data or the inability of social media analytics to cope with enormous and ever-changing datasets. Additional complexities arise from the fact that the available data may lack appropriate quality, accessibility, retrieval, or the ability to be retrieved and combined with other data sources. The relevance and viability of such research may also be constrained by the absence of an established conceptual and theoretical framework in educational scholarship. These intricacies here are all important and require careful consideration.

This study addresses three core questions about MOOC learners on social media:

(1) Which machine learning algorithm optimally identifies MOOC learners within noisy social media data?
(2) What evolving themes and experiences do 'silent learners' express on social media, particularly those absent from formal MOOC platforms?
(3) How can social media discourse inform interventions for underserved MOOC populations?

Literature review

Social media has increasingly become a valuable educational tool, utilised in both MOOCs and traditional classes to motivate student learning and enhance performance and retention rates (Zhao et al., 2020; Zheng et al., 2016). Various studies have explored the use of platforms, like X to support teaching and learning. For instance, Mills (2014) found that pre-service teachers intended to use Twitter as a pedagogical tool in their future classrooms, revealing the platform's potential in educational contexts. Salmon et al. (2015) studied MOOC learners with different levels of Facebook and Twitter usage, analysing their opinions and messages to gain valuable insights into learner perspectives. Additionally, some educators have adopted X as a primary platform to attract learners and disseminate MOOCs (Perifanou & Economides, 2022). Content analysis has also been used to identify topics and themes in posts shared by MOOC learners (Veletsianos, 2017). However, these studies often rely on data from specific MOOC courses or hashtags, and their datasets are limited to Twitter users.

Sina Weibo, often termed 'China's Twitter', is a hybrid platform combining microblogging, multimedia sharing, and real-time public discourse. With over 590 million monthly active users (as of 2024 Q4), it serves as a vital space for educational discussions due to its accessibility, informal tone, and mobile integration. Unlike Western platforms, Weibo operates within China's unique socio-technical ecosystem, where educational content coexists with commercial and institutional voices, offering a rich yet challenging dataset for learner identification. These unique characteristics make Weibo posts an alternative channel for unveilling the often-overlooked aspects of MOOC learning. These authentic digital footprints are particularly valuable, especially when compared to course surveys or questionnaires, which typically have low response rates in MOOCs (Hollister et al., 2022). Analysing such posts can offer researchers and developers valuable insights into learners' experiences, challenges, and difficulties (Zheng et al., 2016). However, research utilising microblogs as a data source in educational studies remains limited.

One challenge in using social media for educational research is accurately identifying actual MOOC learners. As an open platform, anyone can post content, making it difficult to distinguish genuine learners. Previous studies have found that a significant portion of MOOC-related posts are generated by prominent users (e.g. VIP accounts) and MOOC organisations or institutions (Shao et al., 2023). Therefore, the first step in analysing learner discussions is to accurately identify actual MOOC learners. This task is particularly challenging because microblog data tend to be noisier and its users more heterogeneous than those on dedicated MOOC learning platforms like Coursera, edX, and Udacity.

Silent learners in MOOCs represent a significant yet understudied population. While traditional MOOC research has focused on forum participants (who represent only 5–10% of enrollees, according to Ferguson & Clow, 2017), silent learners – those who minimally engage in platform discussions but may express opinions elsewhere – constitute the majority. Duran (2020) revealed that online silence is often purposeful, with learners actively observing and reflecting before participating. Chen et al. (2022) further distinguished between 'silent learning' (content consumption like video watching) and 'active learning' (content contribution), demonstrating that silent behaviours can positively impact learning outcomes.

Our study extends this understanding by examining silent learners who voice their MOOC experiences on social media rather than on course platforms. As Azmat and Ahmad (2022) showed, a lack of interaction can lead to psychological distress, suggesting that silent learners may seek alternative outlets like social media for expression. This aligns with Veletsianos's (2017) finding that 42% of disengaged learners critique courses more candidly on external platforms.

Research method

Microblogging, a widely popular form of social media among young individuals, has increasingly become a hub for educational and teaching activities. This study encompassed a comprehensive analysis of textual content extracted from Weibo posts mentioning 'MOOC' or 'Muke (慕课)' between 2010 and 2022. From an initial collection of 221,336 Weibo posts (2013–2022) mentioning 'MOOC', we employed stratified random sampling (1,800–2,200 posts per month) to ensure temporal representation. Years with fewer than 1,000 posts (2010–2012) were excluded from the analysis.

This study encompasses a heterogeneous array of MOOCs, reflecting the diversity inherent in open online education. The analysed courses span formal higher education programmes (e.g. Peking University's courses on 'Chinese University MOOC') and informal skill-based courses (e.g. vocational training on NetEase Cloud Classroom). This dual scope aligns with recent classifications of MOOCs as hybrid ecosystems bridging academic and lifelong learning contexts. Notably, our keyword-based data collection ('MOOC' or '慕课') intentionally avoided platform-specific limitations (e.g. Coursera vs. domestic platforms) or certification types (free vs. paid credentials).

The inclusivity of various MOOC types is further justified by empirical evidence on learners' diverse motivations. For instance, Veletsianos (2017) demonstrated that MOOC participants range from degree-seeking students to hobbyists seeking personal growth, with 42% of learners engaging in non-certified courses for informal skill development. By capturing this breadth, our analysis transcends narrow platform-centric biases, offering insights applicable to global MOOC ecosystems despite the Weibo-centric dataset.

A. Data preparation

To ensure the integrity of the social media data mining process, it was necessary to clean the Weibo posts by removing emojis, special characters, and other irrelevant elements lacking semantic meaning. Given the unique nature of the Chinese language, which lacks spaces between characters and words, Chinese word segmentation was initially performed. Additionally, as Chinese sentences often contain prepositions, exclamations, modal particles, and verbal expressions reflecting personal habits, these linguistic elements were removed after the word segmentation process to enhance the precision of subsequent text analysis and classification.

Prior to implementing machine learning and deep learning algorithms to identify MOOC learners, the data underwent pre-coding to create a training set for text classification. In this study, approximately 40,000 posts from 2013 to 2022 were manually coded to minimise sampling errors. Research assistants assigned a value of 1 to posts associated with recognised MOOC learners and a value of 0 to posts unrelated to such learners.

The overall distribution of labelled microblog posts between 2013 and 2022 revealed a significant imbalance between learners and non-learners. Specifically, there were 50,443 posts authored by non-learners and 11,742 by learners. The skewed nature of the dataset significantly impacts the performance of binary classification models. Common methods employed to address sample imbalance can be categorised into three types: data-level, algorithm-level, and ensemble solutions (Zhu et al., 2017). Data-level solutions attempt to rebalance the class distribution by resampling the original dataset, including techniques such as random undersampling, random oversampling, and data augmentation. Algorithm-level methods enable the model to prioritise certain classes of samples. Ensemble solutions attempt to enhance the performance of a single classifier by combining multiple classifiers.

In this study, to address the sample imbalance, random undersampling of the non-learner samples was performed. This resulted in a balanced sample set comprising 12,302 non-learner samples and 11,742 learner samples. Subsequently, the sample was divided into training, testing, and validation sets using a ratio of 7:1.5:1.5 for deep learning models and a ratio of 7.5:2.5 for machine learning models.

B. Data analysis: Text classification using machine learning and deep learning

Initial analysis of posts related to ‘MOOC’ revealed that microblogs generate large, diverse, and time-sensitive data. However, discussions on Weibo pertaining to ‘MOOC’ involve not only MOOC learners but also celebrities and learning institutions. Compared to dedicated MOOC learning platforms like Coursera, edX, and Udacity, microblog data tends to be noisier and its users more heterogeneous. Thus, accurately identifying learners on the microblogging platform is crucial for further data-intensive educational research but poses significant challenges.

In this study, we compared the performance of different models to determine the most suitable algorithm for short text classification. We employed a range of machine learning and deep learning methods, considering factors such as data volume, noise level, and the specific characteristics of Weibo text. Machine learning classifiers – including Logistic Regression, Support Vector Machines, Naive Bayes, XGBoost, Random Forest, K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) – were implemented to explore their effectiveness and suitability for Weibo text classification, based on their respective trade-offs and characteristics. Deep learning algorithms, such as TextCNN, TextRNN, and Transformer, are specifically designed to handle the challenges of text classification. TextCNN leverages the feature extraction capabilities of Convolutional Neural Networks (CNNs), often incorporating pre-trained word vector models (e.g. Word2Vec) to enhance contextual information between words and improve classification accuracy. TextRNN, on the other hand, addresses the inherent sequential nature of text by considering the connections between words in a sentence. The Transformer architecture offers a radical departure from traditional models by allowing parallel processing and capturing long-range dependencies between words. By evaluating the performance of these models, we aimed to identify the most effective approach for short text classification in the context of Weibo data. This analysis contributes to the advancement of research on MOOC learner identification and provides valuable insights for utilising social media data in educational studies.

C. Data analysis: Topic model analysis

Utilising the most effective MOOC learner identification algorithm identified in previous analyses, we aimed to extract MOOC learner posts from a substantial sample of 221,336 social media posts. After processing 63,206 text posts were attributed to MOOC learners. To ensure data integrity, duplicate texts were removed, resulting in 40,167 distinct posts authored by genuine MOOC learners. We further refined the dataset by considering only posts with a length of six or more words, yielding 40,147 posts that were subjected to topic model analysis.

For analysing social media data through text clustering, we chose to employ unsupervised machine learning techniques, specifically the Latent Dirichlet Allocation (LDA) topic model. This decision was based on the LDA model’s ability to infer topic distributions within documents, capturing the semantic relationships between words and documents through word co-occurrence and probability within specific semantic contexts. Given the extensive temporal span of our data, we recognised the need to examine the influence of time on discussion topics. Consequently, in addition to the LDA topic model, we employed Dynamic Topic Modeling (DTM) to explore trends in topic evolution over time.

Results

A. Identifying MOOC learners

To evaluate the efficacy of various algorithms in identifying genuine MOOC learners, several machine learning models were compared. The training set comprised 18,033 posts, while the test set consisted of 6,011 posts. Two feature extraction techniques based on TF-IDF and n-gram were employed for text vectorisation in the machine learning models. The accuracy of each model was assessed using the same dataset division and feature processing methods, as presented in Table 1.

The results in Table 1 indicate that the highest accuracy among the machine learning classification on the test set was achieved by the Support Vector Machine (SVM) model using a combination of TF-IDF and n-gram for text vectorisation,with an accuracy of 0.895. Furthermore, the combination of TF-IDF and n-gram consistently outperformed the standalone TF-IDF approach across all models. This underscores the enhanced effectiveness of incorporating n-gram-based feature extraction for this classification task.

The deep learning models also demonstrated strong performance, with TextCNN achieving the highest accuracy of 0.908, establishing it as the top-performing classification algorithm.

Word vector dimensions is a significant parameter that impacts model performance. To investigate this effect comprehensively, we trained model with varying word vector dimensions, (specifically, 100, 150, . . ., 450, and 500) across different model architectures. The corresponding accuracy rates under these diverse dimension sizes are plotted in Figure 1. Although different models achieved peak accuracy at different dimension sizes, the optimal dimensions appeared to be around 150 and 250, which align closely with the length distribution of microblog texts. Notably, for shorter microblog texts, TextCNN exhibits superior performance when compared to TextRNN and Transformer.

Table 1. Comparison of classification accuracy of machine learning algorithm models.

Model	Feature Processing	Accuracy
KNN	TF-IDF	0.765
	TF-IDF+n-gram	0.769
Naive Bayesian	TF-IDF	0.867
	TF-IDF+n-gram	0.876
Logistic Regression	TF-IDF	0.880
	TF-IDF+n-gram	0.891
MLP	TF-IDF	0.863
	TF-IDF+n-gram	0.871
Random Forest	TF-IDF	0.889
	TF-IDF+n-gram	0.893
SVM	TF-IDF	0.887
	TF-IDF+n-gram	**0.895**
XGBoost	TF-IDF	0.884
	TF-IDF+n-gram	0.885

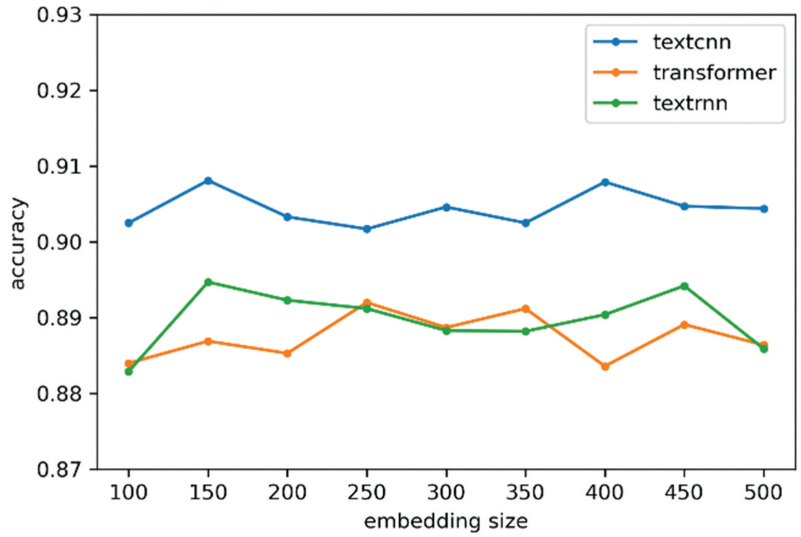

Figure 1. Comparison of the accuracy rate of models based on different word vector dimensions.

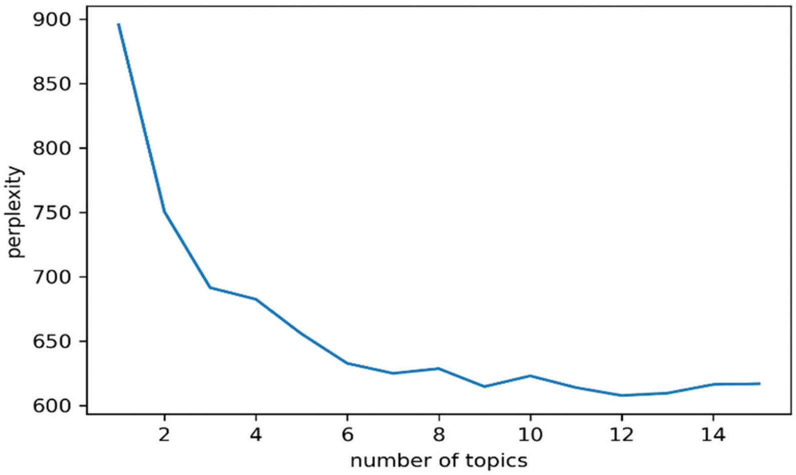

Figure 2. The perplexity of LDA topics.

B. Identifying the themes of MOOC discussions

To determine the optimal number of topics for the LDA model, this study employed a perplexity curve, as illustrated in Figure 2. After analysis, the optimal number of topics was identified as 9, from the choices of 7, 9, and 12.

The application of LDA topic modelling successfully uncovered nine prominent themes that comprehensively encapsulate learner discussions on the platform. These themes – MOOC learning community, platform user experience, learning experience, course recommendation, certificate sharing, learning progress, assignments and exams, course learning, and learning monitoring – provide valuable insights into learners' interests and

Table 2. Topics of MOOCs discussion using LDA.

Overarhing Dimension	Sub-Dimension	Topic	Topic Weight	Keywords (top15)
Open sharing	Information sharing	MOOC learning community	0.108	MOOC, MOOC academy, knowledge, courses, Guokr, learning, affiliated, learning community, certificate, study buddy, classroom, Coursera, Netease, notes, online courses
		Course recommendation	0.083	MOOC, learning, course, currently studying, one course, not bad, participate, together, programmer, usage, DreamWorks, revolution, Android, notes, introduction
	Experience evaluation	Platform using experience	0.131	MOOC, launch, just now, online courses, mobile phone, software, computer, Chinese University MOOC, cannot, classroom, attending class, notes, trash, live broadcast
		Learning experience	0.156	teacher, MOOC, no, not, feeling, attending class, course, problem, things, university, school, PPT, lecture, serious, self-learning
Self-presentation	Self-expression	Certificate sharing	0.081	certificate, MOOC, finish, Chinese University MOOC, complete, course, teaching mode, certified, submitted, start a class, assignments, authentication, friend, brief, regular
		Course learning	0.084	learning, Chinese University MOOC, production, world, postgraduate entrance examination, Tencent meeting, university, MOOC, Harbin Institute of Technology, music, principle, intermediate, financial accounting, management, coaching.
	Self-control	Learning progress	0.147	today, MOOC, tomorrow, no, evening, time, afternoon, feeling, one day, now, hour, every day, morning, hope
		Assignments & exams	0.118	MOOC, assignments, exam, whine, semester, final examination, deadline, grades, not, forgot, know, find, test, missed, now
		Learning monitoring	0.091	MOOC, English, clock-in, ah, assignments, review, learning, words, notes, postgraduate entrance examination, finish, today, reading, calculus, basic

concerns. These nine topics can be grouped into four sub-dimensions (information sharing, experience evaluation, self-expression, and self-control), which are further classified into two overarching dimensions: open sharing and self-presentation, as presented in Table 2.

It is worth noting that our LDA topic modelling findings are consistent with our prior study (J. Zhang et al., 2019), in which we manually coded 4,000 postings to discover the prevailing themes of MOOC learning on Sina Weibo. Within the dimension of open sharing, learners actively engaged in sharing articles and notes within the MOOC learning community, fostering a collaborative learning environment. Additionally, learners extensively discussed their experiences with various MOOC platforms, with a notable emphasis on Chinese University MOOC. Furthermore, they enthusiastically shared their learning experiences with MOOC courses, creating an informative discourse for others. Learners were also eager to recommend courses they were presently undertaking, which primarily focused on computer-related subjects.

However, it is essential to acknowledge that not all discussions were entirely positive. Some learners expressed negative sentiments about their platform experiences and course learning, referring to MOOC platforms as ‘garbage’ and voicing opposition to specific teaching methods, such as reading from PowerPoint slides (PPT). This aspect highlights the diversity of learners' perspectives and the importance of understanding and addressing their concerns to enhance the overall MOOC learning experience.

In the self-presentation dimension, learners predominantly engaged in activities aimed at presenting themselves and their achievements. This involved showcasing the MOOC certificates they had earned, underscoring their accomplishments and credentials. Additionally, learners actively documented their recent learning progress, providing insights into their continuous growth. Furthermore, learners expressed concerns about assignment and exam deadlines, indicating their dedication to meet academic requirements. They also shared detailed accounts of the diverse courses they had studied, spanning various subjects such as music, finance, and management, showcasing their broad educational pursuits. Moreover, learners used the microblogging platform to maintain records of their daily learning activities, illustrating their commitment to continuous learning and self-improvement. These self-presentation endeavours not only enabled learners to project a comprehensive image of their learning journey but also enriched the overall discourse on the microblog platform by fostering a dynamic and interactive learning community.

C. The evolving discussion topics

Dynamic Topic Modeling (DTM) was employed to investigate how discussion topics among MOOC learners evolved from 2013 to 2022. The analysis was divided into three distinct stages; the initial stage (2013–2015), the middle stage (2016–2018), and the recent stage (2019–2022). These divisions allowed us to observe how the discourse changed over the years, revealing valuable insights into shifting trends and interests within the MOOC community. Figure 3. visually depicts the dynamic nature of these discussion topics, highlighting the patterns and variations that emerged during each stage.

During the initial stage (2013–2015), learners primarily centred their discussions on sharing their learning experiences, exchanging notes and articles within the MOOC learning community, sharing insights about MOOC platforms usage, and actively posting their daily learning activities.

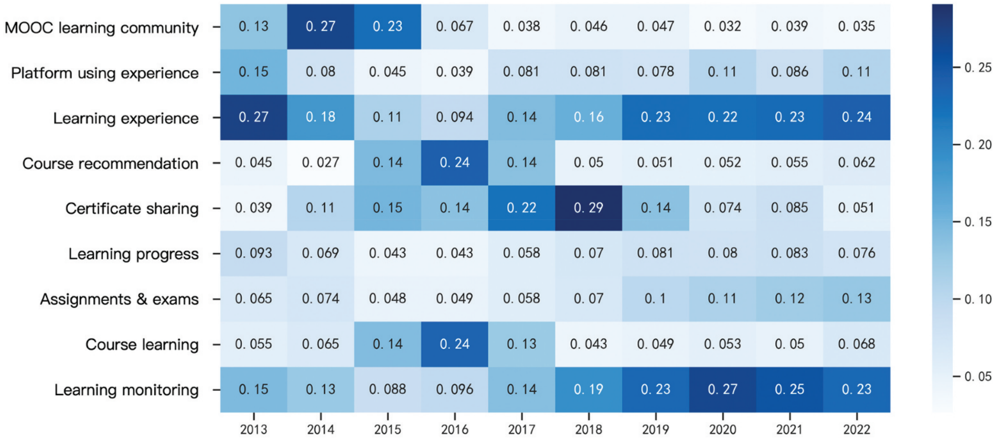

Figure 3. The intensity of text topics varies over time.

In the middle stage (2016–2018), learners increasingly focused on sharing their hard-earned certificates, enthusiastically recommending courses to peers, and posting about the courses they were currently taking. Despite these shifts, the habit of diligently recording their learning activities through 'clock-ins' remained prominent.

In recent years (2019–2022), a discernible transformation in discussions unfolded. The fervour for sharing certificates gradually waned, while discussions about learning experiences and learning clock-ins resurged in popularity. Moreover, conversations related to assignments and exams displayed a steady annual increase. This noteworthy shift is closely associated with the widespread adoption of home-based MOOC learning during the COVID-19 pandemic. As MOOCs assumed a pivotal role in the education landscape, an influx of primary, secondary, and college learners embraced MOOC platforms and actively engaged in discussions. Participants demonstrated an eagerness to share their experiences, provide comments, and initiate numerous discussion threads on topics like learning experiences, clock-ins, platform usage, assignments, and exams on social media platforms like Weibo. This dynamic surge in participation highlights the pivotal role MOOCs played during the pandemic, particularly among a diverse range of learners who become both active participants and enthusiastic contributors to the MOOC community discourse. This transformation reflects how learners adapted to the unique circumstances of the pandemic, underscoring the vital importance of MOOCs as an accessible and inclusive mode of education during challenging times.

Discussion

The current study employs data classification and topic modelling technologies to identify MOOC learners from social media, offering valuable insights into their perceptions of MOOCs. A significant contribution of this research lies in shedding light on silent learners – a population often overlooked in studies relying solely on formal MOOC platform data (Bozkurt et al., 2016). The following discussion explicates how our findings provide clear answers to each research question (RQ).

A. Addressing RQ1: Optimizing MOOC learner classification with advanced algorithms

This section directly responds to RQ1, which sought to identify the optimal machine learning algorithm for discerning genuine MOOC learners within noisy social media data. Our analysis of MOOC-related posts on Weibo identified 63,206 out of 221,336 posts (28.6%) as learner-authored. This ratio is slightly lower than the 35% reported by Veletsianos (2017), a discrepancy likely attributable to methodological differences in data collection; our study used general keywords such as 'MOOC' or '慕课', whereas Veletsianos focused on specific course hashtags, potentially yielding a more concentrated dataset.

Despite these methodological differences, it is evident that only around one-third of the social media posts on platforms like Weibo are made by MOOC learners. The majority are from other sources, including organisations, instructors, MOOC platforms, and academic institutions. This distribution underscores the challenge of analysing social media data in educational research due to the significant 'noise' present. Consequently, the application of robust classification algorithms is paramount. By doing so, educational studies can more effectively leverage these datasets to delve deeper into the experiences and perspectives of MOOC learners, a demographic that has been somewhat under-represented in prior research.

The TextCNN network structure exhibits a clear advantage in short text classification tasks, and the most effective word vector dimensions tend to concentrate around 150 and 250. Compared to RNNs, TextCNN can learn model parameters with a relatively simpler structure and less training data (C. Zhang et al., 2022). However, this finding contrasts with the study by Yang et al. (2021), where a BERT model based on an attention mechanism showed the best performance in classifying the degree of depression expressed in Weibo posts. This discrepancy could be related to the number of transformer encoder blocks used (i.e. the depth of the model). It is noteworthy that machine learning-based text classification algorithms, such as SVM and random forest, achieve an accuracy of approximately 90%, which is comparable to that of deep learning-based text classification algorithms. Therefore, for MOOC learner identification, deep learning neural network algorithms are highly recommended.

While the TextCNN model has demonstrated satisfactory performance in this study, it is noteworthy that deep learning classification models often require a substantial amount of annotated data for effective training. Leveraging large language models, some research endeavours have achieved comparable or even superior text classification capabilities with reduced training data usage, as evidenced in legal cases (Liga & Robaldo, 2023) and educational scenarios (Álvarez-Álvarez & Falcon, 2023). Future research could further investigate whether large-scale models exhibit superior performance in the identification of MOOC learners.

B. Addressing RQ2: Unveiling the hidden voices and experiences of silent learners

Our findings provide a definitive answer to RQ2, revealing the evolving themes and experiences that silent learners express on social media. Notably, the discourse themes of MOOC learners gleaned from social media differ substantially from those portrayed on the course platforms. While forum posts are often directly related to course content (e.g. accessing resources, asking questions), social media posts are predominantly personal and metacognitive. Learners use these platforms to share experiences and feelings, express reflections, and engage in self-supervision, offering an invaluable window into their authentic learning experiences. This difference may stem from learners' distinct perceptions of social media platforms and course forums. For learners, the social nature of social media platforms, coupled with their informal characteristics, allows users to view it as a space for free self-expression.

This rich data allows us to challenge prevailing assumptions about silent learners. First, contrary to portrayal of silent learners as disengaged (Duran, 2020), their vibrant social media activity – particularly around self-monitoring (Topic 9) and emotional venting (Topic 5) – reveals an alternative engagement paradigm (supporting Theme 1 in Azmat and Ahmad (2022) on purposeful silence). Second, while Chen et al. (2022) found silent learning (eg., video watching) enhances performance, our social media data shows these learners also crave informal discussion spaces absent from formal platforms (echoing Duran (2020)'s isolation findings). Third, 'silence as demarcation' theme manifests in our data as learners using social media to discuss taboo topics like instructor criticism which echoes findings in Azmat and Ahmad's research (2022).

These insights suggest silent learners navigate the 'participation paradox': they resist structured forum interactions but actively seek unstructured peer connections for social learning benefits, a dimension of engagement previously hidden from analytics.

C. Addressing RQ3: Informing interventions through the evolution of discourse

Finally, our analysis of evolving topics offers crucial insights for RQ3, guiding how social media discourse can inform interventions for underserved MOOC populations. In this study, learners not only share their learning experiences and offer comments on courses and instructors but also employ the platform for self-regulation and self-motivation. This finding aligns with prior research by Shao et al. (2023) on MOOC learning experiences on social media.

The DTM analysis revealed significant temporal shifts in discussion themes, powerfully contextualised by external events. The resurgence of topics like 'learning experience' and 'learning monitoring', coupled with a steady rise in discussions about 'assignments and exams' during the COVID-19 pandemic, highlights a period of intensified need for self-regulatory strategies and support. This aligns with previous research results (M. Liu et al., 2016; Veletsianos, 2017), where personal feelings and reflections constitute a significant proportion of learners' forum posts. Simultaneously, it aligns with the ongoing trends in MOOC research, where topics such as self-regulated learning (SRL), learner perceptions, and satisfaction are gaining increasing attention (C. Liu et al., 2021). During the period of remote learning amid the pandemic, the importance of learners' autonomous learning abilities and self-regulation skills gained widespread attention among students, parents, teachers, and the general public. The shift towards self-regulated learning discussions during the pandemic highlights the need for platforms to integrate adaptive support mechanisms, such as AI-driven progress tracking and peer accountability features. Such changes could impact the way students learn in and out of school and may also necessitate changes in future teacher education.

Conversely, the peak and subsequent decline in 'certificate sharing' post-2018 provides critical intelligence for motivational design. The phenomenon can be attributed to two main reasons. The first may be the influx of K-12 learners during the pandemic with no certification needs. The second could be influenced by various factors, such as certification policies, leading to a decreasing trend in the demand for certificates among learners over the years. For instance, whether certificates are fee-based can affect students' choices in obtaining them (Littenberg-Tobias et al., 2020). Discussions about certificates in 2019 decreased by half compared to 2018. It is worth noting that the policy of remote online learning had not yet been implemented in 2019. If the decline stems from waning perceived value, MOOC providers must re-evaluate incentive structures. As research indicates certificates can boost persistence and outcomes (Reeves et al., 2017), their diminishing discussion signals a need for more compelling and relevant achievements to sustain engagement. These temporal patterns, therefore, do not merely describe what was discussed but directly point to when and how specific learner support interventions – focusing on either pedagogical support or motivational design – could be most effectively deployed.

Conclusion

Leveraging data science approaches empowers us to study 'authentic' MOOC learners who remain silent on formal MOOC platforms, laying the foundation for further data-intensive educational research. This study directly addressed its three research questions through machine learning and topic modelling.

For RQ1 (optimal algorithm) and RQ2 (silent learners' posts themes), by comparing various algorithms for identifying learners and categorising their post topics, we gained insights into the nature and content of discourse among MOOC participants. To answer RQ3 (informing interventions), DTM showed topics evolved with events like COVID-19 (e.g. resurgent learning monitoring, declining certificate sharing), highlighting shifting learner needs and offering direct insights for designing better support and incentives.

However, there are limitations and directions for future research. While automated text analysis offers convenience and speed, it may overlook certain nuances, suggesting a need to integrate hybrid methodologies for a deeper exploration of content and topics. such as combining social media analytics with large language models to decode learner behaviours at scale. Employing large language models for text classification could enhance the effectiveness of these analyses. Additionally, examining posts from instructors is equally crucial, as it can provide valuable insights into the teaching aspects of MOOCs.

In conclusion, social media, as an essential source of educational big data, presents a more intricate and comprehensive portrayal of the MOOC learner population compared to formal learning platforms. These informal learning platforms, including social media, provide a unique opportunity to explore educational studies through data-intensive approaches, introduce innovative methods and new data sources, rediscover neglected and under-researched learners who are overlooked by traditional educational methods, and ultimately help make MOOCs the pillars of life-long learning in the digital age.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The funding is to support by Beijing Municipal Commission of Education (No. CAGA25051).

Notes on contributors

Yang Liu is Associate Professor at Institute of Education, China University of Geosciences in Beijing. She took her PhD in Cognitive Science in Education from Teachers College, Columbia University. Her research interests include embodied cognition, big data in education, and education policy. She is also a long term learner and researcher on online education since 2012 when MOOCs started booming on the Internet.

Jingjing Zhang is a Professor of Educational Technology and Associate Dean at the Faculty of Education, Beijing Normal University. Her research focuses on applying data mining techniques to examine human relationships and online activities, with a particular emphasis on the learning sciences. She holds both a Ph.D. and an MSc from the University of Oxford. Prior to joining BNU, she trained at the OECD in Paris and completed an internship at the United Nations Headquarters in New York.

Yehong Yang is a Ph.D. candidate in the Faculty of Education at Beijing Normal University. Her research focuses on artificial intelligence in education (AIED), learning analytics, and blended learning environments.

Xutong Tao holds a Master's degree in Software Engineering from Beijing Normal University and is currently a research scientist in large language models at Baidu.

References

Álvarez-Álvarez, C., & Falcon, S. (2023). Students' preferences with university teaching practices: Analysis of testimonials with artificial intelligence. *Educational Technology Research & Development*, *71*(4), 1709–1724. https://doi.org/10.1007/s11423-023-10239-8

Ansari, J. A. N., & Khan, N. A. (2020). Exploring the role of social media in collaborative learning the new domain of learning. *Smart Learning Environments*, *7*(1), 9. https://doi.org/10.1186/s40561-020-00118-7

Azmat, M., & Ahmad, A. (2022). Lack of social interaction in online classes during COVID-19. *Journal of Materials and Environmental Science*, *13*(2), 185–196.

Barthakur, A., Kovanovic, V., Joksimovic, S., Siemens, G., Richey, M., & Dawson, S. (2021). Assessing program-level learning strategies in MOOCs. *Computers in human behavior*, 117, 106674.

Bozkurt, A., Honeychurch, S., Caınes, A., Balı, M., Koutropoulos, A., & Cormıer, D. (2016). Community tracking in a cMOOC and nomadic learner behavior identification on a connectivist rhizomatic learning network. *Turkish Online Journal of Distance Education*, https://doi.org/10.17718/tojde.09231

Chen, L., Zou, C., Geng, S., & Du, Z. (2022). The influence of students' online silent and active participation on learning performance and experience. *Proceeding of WHICEB*, *2022*(90).

Costello, E., Brown, M., Nair, B., Nic Giolla Mhichíl, M., Zhang, J., & Lynn, T. (2017, May). # Mooc friends and followers: An analysis of Twitter hashtag networks. *5th European MOOCs Stakeholders Summit, EMOOCs 2017*, May 22-26, 2017. Madrid, Spain eds. Kloos, Carlos Delgado, Jermann, Patrick, Pérez-Sanagustín, Mar, Seaton, Daniel T., White, Su (pp. 170–175). Springer International Publishing.

Duran, L. (2020). Distance learners' experiences of silence online: A phenomenological inquiry. *International Review of Research in Open & Distributed Learning*, *21*(1), 82–98. https://doi.org/10.19173/irrodl.v20i5.4538

Fan, Y., Jovanović, J., Saint, J., Jiang, Y., Wang, Q., & Gašević, D. (2022). Revealing the regulation of learning strategies of MOOC retakers: A learning analytic study. *Computers & education* , 178, 104404, https://doi.org/10.1016/j.compedu.2021.104404.

Ferguson, R., & Clow, D. (2017, March). Where is the evidence? A call to action for learning analytics. *LAK '17: Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, March 13 - 17, 2017. Association for Computing Machinery. Vancouver British Columbia Canada (pp. 56–65).

Hollister, B., Nair, P., Hill-Lindsay, S., & Chukoskie, L. (2022, May). Engagement in online learning: Student attitudes and behavior during COVID-19. In *Frontiers in education* (Vol. 7, p. 851019). Frontiers Media SA.

Khalil, M., & Belokrys, G. (2022). What does Twitter say about self-regulated learning? Mapping tweets from 2011 to 2021. *Frontiers in psychology*, 13, 820813.

Liga, D., & Robaldo, L. (2023). Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review*, *51*, 105864. https://doi.org/10.1016/j.clsr.2023.105864

Littenberg-Tobias, J., Ruipérez-Valiente, J. A., & Reich, J. (2020). Studying learner behavior in online courses with free-certificate coupons: Results from two case studies. *International Review of Research in Open & Distributed Learning*, *21*(1), 1–22. https://doi.org/10.19173/irrodl.v20i5.4564

Liu, C., Zou, D., Chen, X., Xie, H., & Chan, W. H. (2021). A bibliometric review on latent topics and trends of the empirical MOOC literature (2008–2019). *Asia Pacific Education Review*, *22*(3), 515–534. https://doi.org/10.1007/s12564-021-09692-y

Liu, M., McKelroy, E., Kang, J., Harron, J., & Liu, S. (2016). Examining the use of Facebook and Twitter as an additional social space in a MOOC. *The American Journal of Distance Education*, *30*(1), 14–26. https://doi.org/10.1080/08923647.2016.1120584

Mills, M. (2014). *Effect of faculty member's use of Twitter as informal professional development during a preservice teacher internship. Contemporary issues in technology and teacher education*. https://www.semanticscholar.org/paper/Effect-of-Faculty-Member's-Use-of-Twitter-as-During-Mills/01ff32db7b3a0114bbc02f4b5d34118dccfa6bdd

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Applied Sciences*, *11*(18), 8438. https://doi.org/10.3390/app11188438

Perifanou, M., & Economides, A. A. (2022). The landscape of MOOC platforms worldwide. *International Review of Research in Open & Distributed Learning*, *23*(3), 104–133. https://doi.org/10.19173/irrodl.v23i3.6294

Reeves, T. D., Tawfik, A. A., Msilu, F., & Simsek, I. (2017). What's in it for me? Incentives, learning, and completion in massive open online courses. *Journal of Research on Technology in Education*, *49*(3–4), 245–259. https://doi.org/10.1080/15391523.2017.1358680

Salmon, G., Ross, B., Pechenkina, E., & Chase, A.-M. (2015). The space for social media in structured online learning. *Research in Learning Technology*, *23*(1), 28507–28514. https://doi.org/10.3402/rlt.v23.28507

Shao, Y., Zhang, J., Costello, E., & Brown, M. (2023). Public perceptions towards MOOCs on social media: An alternative perspective to understand personal learning experiences of MOOCs. *Interactive Learning Environments*, *31*(2), 670–682. https://doi.org/10.1080/10494820.2020.1799413

Shen, C., & Kuo, C.-J. (2015). Learning in massive open online courses: Evidence from social media mining. *Computers in Human Behavior*, *51*, 568–577. https://doi.org/10.1016/j.chb.2015.02.066

Theophilou, E., Hernández-Leo, D., & Gómez, V. (2024). Gender-based learning and behavioural differences in an educational social media platform. *Journal of Computer Assisted Learning*, *40*(6), 2544–2557. https://doi.org/10.1111/jcal.12927

Veletsianos, G. (2017). Toward a generalizable understanding of Twitter and social media use across MOOCs: Who participates on MOOC hashtags and in what ways? *Journal of Computing in Higher Education*, *29*(1), 65–80. https://doi.org/10.1007/s12528-017-9131-7

Wei, X., Saab, N., & Admiraal, W. (2023). Do learners share the same perceived learning outcomes in MOOCs? Identifying the role of motivation, perceived learning support, learning engagement, and self-regulated learning strategies. *The internet and higher education*, 56, 100880.

Yang, T., Li, F., Ji, D., Liang, X., Xie, T., Tian, S., Li, B., & Liang, P. (2021). Fine-grained depression analysis based on Chinese micro-blog reviews. *Information Processing & Management*, *58*(6), 102681. https://doi.org/10.1016/j.ipm.2021.102681

Zhang, C., Guo, R., Ma, X., Kuai, X., & He, B. (2022). W-textcnn: A TextCNN model with weighted word embeddings for Chinese address pattern classification. *Computers, Environment and Urban Systems*, *95*, 101819. https://doi.org/10.1016/j.compenvurbsys.2022.101819

Zhang, J., Perris, K., Zheng, Q., & Chen, L. (2015). Public response to "the MOOC movement" in China: Examining the time series of microblogging. *International Review of Research in Open & Distributed Learning*, *16*(5), 144–160. https://doi.org/10.19173/irrodl.v16i5.2244

Zhang, J., Wang, X., Shen, L., & Jiang, L. (2019). Mooc learning characteristics from the perspective of Weibo: Autonomy, sociability, diversity, and openness. *Distance education in China*, 11, 38–47 (in Chinese).

Zhao, Y., Wang, A., & Sun, Y. (2020). Technological environment, virtual experience, and MOOC continuance: A stimulus-organism-response perspective. *Computers and Education*, *144*, 103721. https://doi.org/10.1016/j.compedu.2019.103721

Zheng, S., Han, K., Rosson, M. B., & Carroll, J. M. (2016). The role of social media in MOOCs: How to use social media to enhance student retention. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale* 419–428. https://doi.org/10.1145/2876034.2876047

Zhu, B., Baesens, B., & Vanden Broucke, S. K. L. M. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*, *408*, 84–99. https://doi.org/10.1016/j.ins.2017.04.015