


Article

Power to the Learner: Towards Human-Intuitive and Integrative Recommendations with Open Educational Resources

Sahan Bulathwela * , María Pérez-Ortiz, Emine Yilmaz and John Shawe-Taylor

Centre for Artificial Intelligence, University College London, London WC1V 6BH, UK

* Correspondence: m.bulathwela@ucl.ac.uk

Abstract: Educational recommenders have received much less attention in comparison with e-commerce- and entertainment-related recommenders, even though efficient intelligent tutors could have potential to improve learning gains and enable advances in education that are essential to achieving the world's sustainability agenda. Through this work, we make foundational advances towards building a state-aware, integrative educational recommender. The proposed recommender accounts for the learners' interests and knowledge at the same time as content novelty and popularity, with the end goal of improving predictions of learner engagement in a lifelong-learning educational video platform. Towards achieving this goal, we (i) formulate and evaluate multiple probabilistic graphical models to capture learner interest; (ii) identify and experiment with multiple probabilistic and ensemble approaches to combine interest, novelty, and knowledge representations together; and (iii) identify and experiment with different hybrid recommender approaches to fuse population-based engagement prediction to address the cold-start problem, i.e., the scarcity of data in the early stages of a user session, a common challenge in recommendation systems. Our experiments with an in-the-wild interaction dataset of more than 20,000 learners show clear performance advantages by integrating content popularity, learner interest, novelty, and knowledge aspects in an informational recommender system, while preserving scalability. Our recommendation system integrates a human-intuitive representation at its core, and we argue that this transparency will prove important in efforts to give agency to the learner in interacting, collaborating, and governing their own educational algorithms.

Keywords: open education; recommendation systems; lifelong e-learning; state-based learner modelling; Sustainable Development Goal 4



Citation: Bulathwela, S.; Pérez-Ortiz, M.; Yilmaz, E.; Shawe-Taylor, J. Power to the Learner: Towards Human-Intuitive and Integrative Recommendations with Open Educational Resources. *Sustainability* **2022**, *14*, 11682. <https://doi.org/10.3390/su141811682>

Academic Editor: Eddie W.L. Cheng

Received: 15 June 2022

Accepted: 9 September 2022

Published: 17 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Equitable quality education is one of the Sustainable Development Goals (SDG). Education (SDG 4) is specifically critical to achieving other SDGs, such as eradicating poverty, improving health and well-being, and providing decent work and economic growth opportunities while passively contributing to many other SDGs (e.g., improving industry, innovation and infrastructure, developing sustainable cities, peace, justice and strong institutions, etc.). However, delivering quality education is a complex challenge at multiple scales and scopes (pedagogically, operationally, and technologically [1]). The world population is very diverse, and education ultimately should support human individuality, allowing learners to work towards their own goals and realise their own dreams at their own pace.

Technology, coupled with recent developments in open education, provides the possibility of personalising education at scale. However, the technological framework can be critical in terms of whether that technology empowers the user or achieves the complete opposite (e.g., does not supporting a diverse population of learners or nudges the user in directions that are not of interest or are pedagogically unsuitable for them).

We envision empowering personalised learning at scale with a learning assistant that supports the user in their learning paths, giving them agency and control, in addition to

access to a plethora of pedagogically and culturally diverse educational resources. We argue that for this technology to be empowering, the learner needs to be able to interact with the notions and perceptions of their own learning models, as well as contest and govern them to suit their own varied educational needs, empowering them with a sort of technological learning “exoskeleton”. Crucial technical components necessary to allow such interaction and collaboration are the transparency and interpretability of the system’s variables (e.g., learner interests). This work focuses on addressing these challenges by developing an artificial intelligence (AI) system that can: (i) infer multiple-user latent variables, such as user interests/goals or knowledge/skills; (ii) create a human-intuitive representation that can be understood, queried, and modified by the user; (iii) understand the topical content of learning materials; (iv) respect population diversity and users’ privacy; and (v) be data-efficient while addressing common recommendation system challenges, such as the cold-start problem [2,3]. We see this as a first step towards the ambitious goal of building AI systems that enable empowering personalised lifelong learning.

Developing user models that can to some extent capture the factors that affect user engagement is foundational to building effective recommendation systems in education [2], as well as for many other applications related to knowledge management and tracing. A recommender attempts to rank items in a collection in order to personalise the individual experience. Educational recommenders achieve the same goal by ranking relevant learning resources differently for individual learners, leading to personalised learning experiences. Traditionally, the user modelling and recommender system research community has focused on modelling different user factors in isolation, such as user interests or knowledge states. However, plenty remains to be performed if we are to build user models that are truly integrative, simultaneously modelling and combining relevant latent user variables. Such models could potentially help us (and the learners) to better understand individualised learning preferences and goals, with the objective of providing the most fulfilling learning path.

Our work consists of two core contributions: (1) a novel online learner model that can infer learner interests based on implicit user interactions (we call it TrueLearn Interest); (2) multiple integrative models that use content popularity, learner interest, novelty, and knowledge to make engagement predictions. The proposed TrueLearn INK and TrueLearn PINK models are online, transparent user modelling algorithms that are also integrative (in the sense that they can model multiple factors that influence learner engagement). Our experiments with a large dataset of more than 20,000 learners indicate that our proposals, TrueLearn Interest, INK, and PINK, can significantly outperform competitive baselines in predicting in-the-wild learner engagement with educational videos, while still utilising multiple classification metrics currently used to evaluate performance.

Although much remains to be performed (on technical, social, and pedagogical fronts) for these personalised learning approaches to start producing practical results, the human-intuitive user and content representation, along with the data/computational efficiency of the TrueLearn model family, makes it a very strong candidate for personalising the learning of masses of informal learners over long periods of time. This paves the way towards building a promising AI companion for lifelong learning that relies on freely available knowledge, such as open educational resources and Wikipedia [1]. All these features and the performance evaluation strongly suggest that such a model family can become the foundation of low-cost, effective, equitable, and quality learning, which will address SDG 4 while at the same time actively and passively benefiting the success of other SDGs.

2. Related Work

The learning analytics and educational data mining (EDM) communities that build intelligent tutoring systems (ITS) and educational recommendation systems (EdRecSys) [4] have developed novel approaches that focus on capturing and exploiting the context of the learner. The contextual factors, such as the knowledge state [5–8], degree of novelty of informational content [9,10], interests [11] and learning goals [12], are crucial variables

that are utilised in developing effective personalised e-learning systems. To build a truly integrative EdRecSys that understands the interplay of these factors, we need to incorporate all of these factors in our recommender model. Figure 1, from our prior work [2], shows a suitable setting to model this interplay. Our primary aim through this work is to build and test the usefulness of a viable algorithm that integrates all these factors. When building EdRecSys, interpretability is crucial as it provides the transparency needed to trigger the trust of a learner [13,14]. Recent studies have further showed that connecting AI recommendations with human-intuitive explanations and visualisations leads to an increase in trust and the perceived sophistication and accuracy of AI-based learning systems [15]. With innovative movements such as open educational resources (OER) [16] and massively open online courses (MOOC) [17] on the rise, our proposals to exploit this massive collection of learning materials should work towards scalable systems that can accommodate populations of learners over a lifetime (e.g., in informal/lifelong learning scenarios). Only through such systems can we truly realise quality education for all (SDG 4). Our previous experiments to tackle this problem have shown promising evidence that building high quality recommenders while retaining scalability and transparency is feasible [10]. The TrueLearn INK and TrueLearn PINK models make further progress in this direction, and resemble Figure 1. They integrate popularity, interest, novelty, and knowledge in an online and transparent learner model.

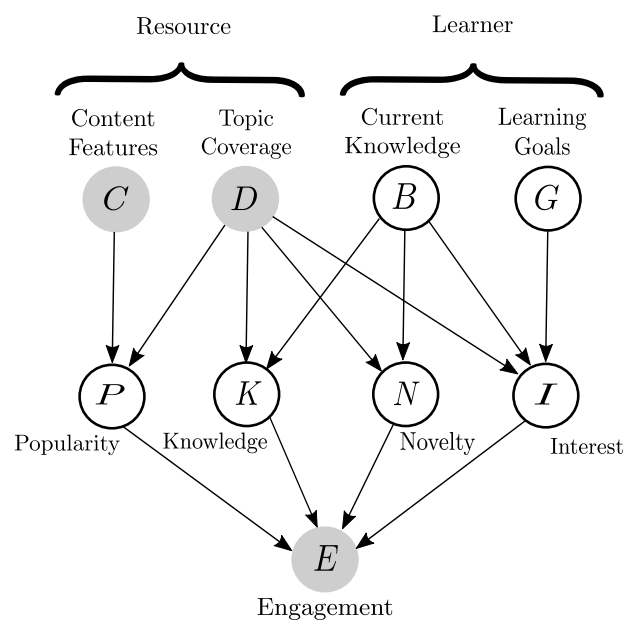


Figure 1. Graphical model of learner engagement that incorporates the drivers of learner engagement, including content (P)opularity, Learner (K)nowledge, (N)ovelty and (I)nterests.

2.1. Opportunities and Challenges in Scaling Personalised Education

The last decade has seen a favourable increase in interest in the research community towards pushing the frontiers of e-learning, with an increased emphasis on the need to address diverse learner backgrounds and goals using large collections of educational materials. In a world populated by MOOCs, many powerful educational resources that can become enablers of our future generations are gated behind licensing walls that limit their reach and impact. OERs provide an excellent remedy for this situation by relaxing copyright restrictions, leading to mass adoption among both learning resource creators [18] and consumers [19]. Currently, new AI-enabled tools that discover and index such collections are emerging [20,21], and these solutions are addressing the vital need for scalable content understanding that can support large OER collections. With the ability to harness these AI-powered tools within our grasp, we should design these openly accessible tools in such a way that we avoid increasing the disparities in terms of access to quality education [1]. In

a landscape where high-quality educational resources are available in abundance, having scalable educational recommenders that can connect distinct learning materials to individuals across languages, cultures, and content modalities over a lifelong journey of learning is a game changer [22].

2.2. Scalable Content Representation

While acknowledging that scalable content-understanding tools are surfacing, it is important to understand their usefulness to EdRecSys. Amongst many approaches, topic-, keyword-, and concept-based approaches have become popular research directions in the context of informational recommendation systems, such as news [23,24], social media [25,26], and educational content [10,27].

From their inception, the ITS and EdRecSys communities have heavily relied on manually labelling the knowledge components (KCs, i.e., knowledge topics covered) in a learning material or exercise [28,29]. However, this is not scalable in practice [5,30]. Since then, unsupervised learning techniques, such as latent Dirichlet allocation (LDA) [31] and entity linking, have shown great promise towards this goal [32]. While unsupervised models, such as LDA, solve the costly human annotation challenge, they entail complexities in hyperparameter tuning and do not guarantee identifying topics that are humanly intuitive. Entity linking handles the explainability of topics by using concepts from a knowledge base, such as Wikipedia, as the topic taxonomy. Multiple models have successfully built content representations using entity linking to leverage informational [33] and educational [34] recommenders, while recent models such as TrueLearn Novel demonstrated very good predictive performance with educational recommendations [10]. Due to this potential of entity linking, we use entity linking for content representation.

Fragments of Content

Finding the relevance of parts of documents rather than entire documents, be it in text [35], audio podcasts [36], or video [21,37], has become a much more prominent research area in recent years. Learning needs may be granular enough for a case where a fragment of a resource is sufficient. Breaking informational videos into fragments has also shown promise in efficient previewing [21,38] and enabling the non-linear consumption of videos [39]. Our prior proposal, the TrueLearn Novel [10] model, demonstrates the potential of utilising fragment-wise recommendation in educational recommenders using the PEEK dataset [40]. This allows the e-learning system to have video fragments that contain a satisfactory amount of knowledge while ensuring that the video fragment length is good for retaining viewer engagement [41]. As we extend the TrueLearn Novel in this work, our new proposals address the fragment-wise engagement prediction task.

2.3. Learner Interest, Novelty, Knowledge, and Content Popularity

The majority of work in the EdRecSys domain revolves around using learner context [8–10] to predict learner engagement. When it comes to the learner context, modelling the interests and personal goals of individual learners plays a key role. This task is an active research area in information retrieval and in the recommendation systems domain. Interest models in our prior work mainly use two approaches for exploiting explicit (especially in e-commerce and entertainment domains [42,43]) and implicit feedback (mainly in social media [44]) as target labels. Using the features of the content that users interact with in order to build individualised profiles is a popular approach when it comes to modelling interest. As a reliable content-based feature, keywords/concepts/entities/topics are widely used. These techniques in unison are identified as concept-based approaches [44]. Many concept-based approaches count the number of times a user interacts with different concepts based on their interaction with materials that contain these concepts. These frequency-based methods build a representation of user interest over time and then use a selected similarity metric (e.g., Cosine or Jaccard similarity) to estimate the most relevant content for the user [27,40,45]. Previously, frequency-based user models [46] built for

educational information retrieval have utilised probabilistic Bayesian modelling inspired by knowledge tracing [30]. While our proposed interest model utilises the frequency of concept occurrence to build a user model, Syed and Collins-Thompson tackle information search (where a query dictates the information need), a different task in comparison with the content recommendation task that we focus on. Other probabilistic approaches proposed for interest modelling rely on unsupervised topic detection models, such as LDA, to discover concepts and model interests in these concepts (e.g., with content merging [47] in a streaming setting [48], etc.). As discussed in Section 2.2, such approaches have their disadvantages and are not relevant to this work as we use entity linking to link with Wikipedia concepts. In recent years, there have also been proposals that utilise deep learning to learn from user interaction sequences in order to make content recommendations. This has led to recurrent neural network models being proposed for EdRecSys [49,50]. However, deep learning approaches lack the human-intuitive representations that we believe to be a key characteristic of the ideal educational recommender [2].

The recent work we identified to be most relevant to the proposed interest model uses concept-based user modelling to recommend MOOCs to learners [34]. In their approach, they consider the user session to be a document in which the topics they visit over time are terms (words). They compute the term frequency (TF) for each user over time to build a user profile. To predict compatibility with potential future content at time t , the user's TF-based topic profile up to time point $t - 1$ is used. The engagement is predicted by measuring the similarity between the user profile and the content. This similarity measure can be used to rank recommendations. Piao and Breslin further extend this method using term-frequency-inverse document frequency (TFIDF)-based profiles, which penalise general topics that occur abundantly and challenge the task of discriminating between users. However, we go beyond naïvely counting concept occurrences to formulate a novel Bayesian model that can capture the relationship between engagement and latent user interest in order to predict the compatibility between the user and the prospective educational video. This sophistication is added using extremely computationally efficient algorithms to ensure the scalability of the models. A scalable, human-intuitive interest representation of the user is created as a result.

The receptiveness of a learner to a specific learning resource depends on their current knowledge state as much as their interests and goals. Knowledge tracing (KT) [5] and item response theory (IRT) [51] are the two main approaches used by the EDM and ITS communities when modelling the knowledge/skills of learners. In the context of KT, Bayesian modelling of the learner interaction journey in the form of a hidden Markov model has been the dominant idea. This approach allows one to recover a rich, interpretable representation of the learner knowledge (i.e., skill mastery) while building a probabilistic model that operates under human-intuitive assumptions [30]. However, the widely available implementations of KT use expectation maximisation to learn, incurring significant computational costs and limiting its use in large-scale applications. Bishop et al. [52] recently demonstrated that KT can be trained online, allowing it to scale better. The interest tracing model described in Section 3.4.1 in this work is based on this idea. Advancements in using machine learning methods (e.g., knowledge tracing machines [6]) and neural KT [7,8]) have improved the predictive performance in certain datasets; however, their superiority is being rigorously questioned in more insightful studies [53,54]. Additionally, these are data-hungry, complex models and lack interpretability. More relevant proposals to our work utilise variants of IRT, such as the Elo [55] and TrueLearn Novel [10] algorithms. The TrueLearn Novel algorithm models novelty in addition to the learner-knowledge state, producing leading results in video engagement prediction tasks using online learning. The utilisation of online Gaussian expectation propagation while maintaining a Wikipedia-based user model retains both scalability and human-intuitiveness, which we maintain in our resultant models. When building TrueLearn INK and PINK models, we use the TrueLearn Novel model to model the knowledge acquisition and novelty assumptions in learning.

Although the connection between contextual learner engagement (inferred from learner interests and knowledge) and learning gains have been explored by many [56,57], we need to understand the role that context-agnostic engagement has to play in this equation; that is, what attributes of educational materials lead to a learner population's engagement with it (regardless of individual learner contexts)? Prior work has identified verticals, such as freshness, presentation, authority, understandability, and topic coverage, as factors that contribute towards population-based engagement, while machine learning models based on these features can be built to predict population-based engagement [58]. Our prior work also indicated that such models were critical in addressing both user and item cold-start problems [3,58]. In this work, we train such a population-based predictor to incorporate with TrueLearn PINK to help the personalisation model address the user cold-start problem.

2.4. Combining Predictions

As argued in Section 2.3, learner engagement is affected by a multitude of different factors. While systems modelling these factors in isolation can be used successfully, the aim is to combine the different hypotheses together to build a more integrative recommender (as per Figure 1). Such a system is capable of accounting for multiple factors associated with the learner. Therefore, combining models is a sensible approach towards utilising all the available hypotheses together (popularity, knowledge, novelty, and interest). There are many ways to combine prediction models [59]. Using probabilistic modelling to combine predictions is one of the simpler ways of doing this [60]. Ensemble techniques (such as Bayesian averaging, boosting, bagging, or stacking) have also shown promise in robustly combining classifiers [61]. Amongst ensemble methods, using a meta-learner (a.k.a stacking) is an approach that allows training a meta-model that learns to weigh the different hypotheses that are being combined. Stacking has been used successfully (with a reported 10% performance improvement) in the learner modelling domain [62,63]. We also experiment with multiple online meta-learners, including a novel model, to create a transparent, scalable, yet integrative, learner model. Another type of combination used in hybrid recommender systems is switching, where the hybrid recommender switches between different models for different predictions, something that we also experiment with. Many recommendation systems utilise switching to start recommendations with a population-based model and switch to a personalised model addressing the cold-start problem [64]. In addition to the meta-learner approach, we experiment with switching to formulate a different version of the TrueLearn PINK model.

3. Integrative and Personalised Educational Recommendations

With our familiarity of the relevant recent works, we outline the experimental methodology in this section. Modelling and predicting user engagement using watch time with videos has been studied in both population-based [41,65] and personalised [40,66] contexts. This paper attempts to build a novel learner model that can predict learner engagement with video lectures using factors such as content popularity, learner interests, novelty, and knowledge, which are outlined in Section 2. Towards this goal, we begin by outlining the problem setting and identifying the different components that need to be developed and combined. We formulate a Bayesian model that captures the assumptions relating to learner interests as the first step. Then, we investigate how we can integrate popularity, knowledge, novelty, and interest models into one model using suitable methods of combination.

3.1. Problem Setting

In this scenario, we focus on modelling where a learner ℓ in a learner population L interacts with a series of educational resources $S_\ell \subset \{r_1, \dots, r_R\}$, where r_x are fragments/parts of different educational videos. The watch interactions happen over a period of T time steps, with R being the total number of resources in the system. In this system with a total N unique KCs, resource r_x is characterised by a set of top KCs or topics

$K_{r_x} \subset \{1, \dots, N\}$. We assume the presence i_{r_x} of KC in resource r_x and that the degree d_{r_x} of KC coverage in the resource is observable. Figure 2 illustrates the problem setting we aim to model.

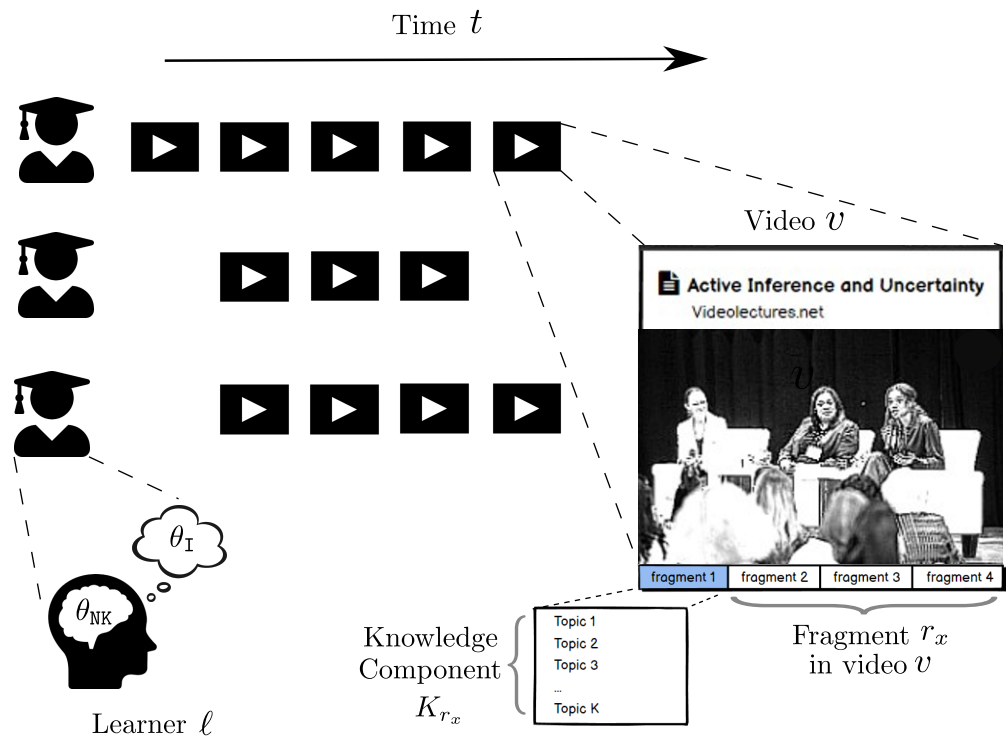


Figure 2. Visual illustration of the problem setting, where learner ℓ with knowledge (that allows them to tackle novel content) θ_{NK} and interests θ_I is watching fragments of educational videos r_x containing different knowledge components K_{r_x} over time t .

The key idea is to model the probability of engagement $e_{\ell,r_x}^t \in \{1, -1\}$ between learner ℓ and resource r_x at time t as a function of the learner interest $\theta_{\ell_I}^t$ and knowledge $\theta_{\ell_{NK}}^t$ based on the top KCs covered K_{r_x} using their presence i_{r_x} , in addition to the depth of topic coverage d_{r_x} .

According to Bayes rule, the posterior distributions are proportional to:

$$P(\theta_{\ell_I}^t | e_{\ell,r_x}^t, K_{r_x}, i_{r_x}) \propto P(e_{\ell,r_x}^t | \theta_{\ell_I}^t, K_{r_x}, i_{r_x}) \cdot P(\theta_{\ell_I}^t) \tag{1}$$

$$P(\theta_{\ell_{NK}}^t | e_{\ell,r_x}^t, K_{r_x}, d_{r_x}) \propto P(e_{\ell,r_x}^t | \theta_{\ell_{NK}}^t, K_{r_x}, d_{r_x}) \cdot P(\theta_{\ell_{NK}}^t) \tag{2}$$

The underlying assumptions for how learner interest, knowledge, and novelty relate to learner engagement are graphically illustrated in Figure 3. Hypothesis (i) represents the common assumption in interest-based informational recommendation systems, namely that a user will engage with an item if there is enough interest towards the topics present in the resource [67]. Hypothesis (ii) is the underlying assumption of TrueLearn Novel [10], which accounts for the knowledge state of the user and the novelty element using ε as the engagement margin. In the subsequent steps, we formulate Bayesian graphical models that resemble the interest hypothesis in Section 3.4; furthermore, we use the TrueLearn Novel model [10] to model the knowledge and novelty hypothesis.

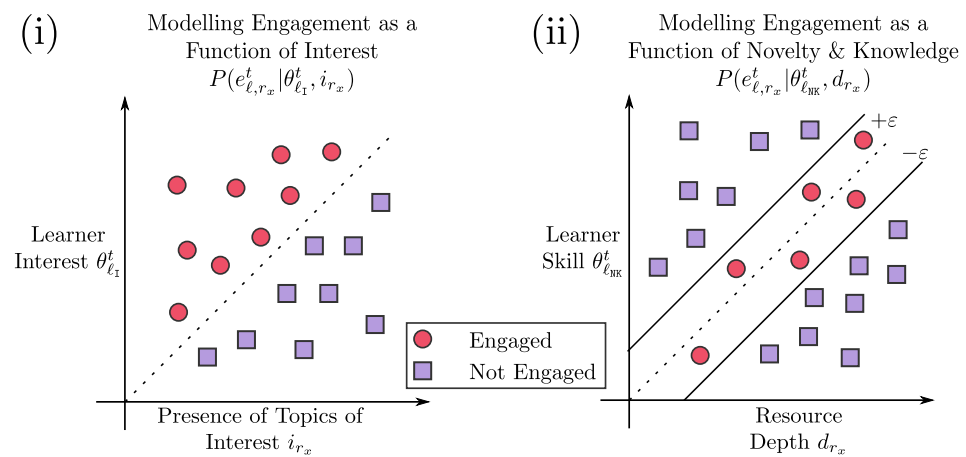


Figure 3. Graphical illustration of the assumptions made when modelling learner's (i) interest (I), and (ii) novelty and knowledge (NK). TrueLearn Interest model proposed in this work tests hypothesis (i). TrueLearn Novel model used in TrueLearn INK and PINK uses hypothesis (ii).

3.2. Data

Two publicly available datasets are used in the experiments and they are based on our prior work. As both datasets are drawn from the VideoLectures.Net repository, the videos overlap. We primarily use the recently published PEEK dataset [40] for the final evaluation as this dataset is the only publicly available dataset perfectly capturing the problem setting stated in Section 3.1. The dataset contains over 20,000 in-the-wild learners consuming fragments of over 10,000 unique educational video lectures. Each lecture fragment in the dataset is also annotated with the top five Wikipedia concepts retrieved via entity linking. The PEEK dataset only includes learners with at least five recorded sessions. The dataset consists of learner sessions where the majority of the sessions are relatively short; however, there is a proportion of user sessions that are comparatively lengthy. The majority of lectures that are found in the PEEK dataset focus on data science/machine learning and computer science, although lectures relating to other subjects are also present.

The VLE dataset [3] was used to train the context-agnostic engagement predictor. A population-based engagement prediction model has to be trained in order to estimate the engagement when combining the population prior. To avoid data leakage, the set of video lectures in the test data of PEEK dataset was identified beforehand and removed from the training data of the context-agnostic model.

3.3. Baseline Models

To measure the comparative performance of the proposed models, we devised several baselines. We used three content-based similarity baselines, namely the (i) Cosine, (ii) Jaccard_C, and (iii) Jaccard_U models, along with multi-skill knowledge tracing and TrueLearn Novel models, which are proposed in prior work [40]. As per Section 2, we identified additional frequency-based learner models proposed in the concept-based interest modelling domain as suitable baselines. Hence, we included the TF- and TFIDF-based user interest models [34] in our baseline set. The PEEK dataset already contains topic depth d_{r_x} . Therefore, we experimented with two versions of the TF model, where (iv) the TF(Binary) model uses the presence of a concept in binary form (i_{r_x}) and (v) the TF(Cosine) uses the coverage of a concept (d_{r_x}) estimated using the Cosine similarity between the video fragment transcript and Wikipedia page of the concept [40]. Finally, (vi) the TFIDF(Cosine) model normalised the concept counts with inverse document frequency.

3.4. Learner Interest

When modelling interest, we hypothesised that a learner starts watching a video fragment because they are interested in the topics included in that fragment. Regardless of whether they engage positively (watching a significant fraction of the video fragment),

choosing to start watching can be considered as a signal of interest in the topics covered. This leads us to use both positive and negative engagement events as positive interest signals. Although a proxy of KC coverage (d_{r_x}) is available to the system [10], we need to assume that the learner is only aware of the fact that a specific topic is present in the video fragment when committing to start watching. Therefore we used binary features i_{r_x} to indicate that a subset of KCs, K_{r_x} were present in a video fragment when modelling interest. To model the assumptions illustrated in Figure 3i, we formulated two online learning models that used density filtering (posterior of $\theta_{\ell_I}^{t-1}$ becomes prior for event at time t) to model learner interest, $\theta_{\ell_I}^t$. The two proposed models are described below.

3.4.1. Interest Tracing Model

As the initial proposal for modelling interest, we formulated this model. This is a user model where the interest parameter for each user is modelled as a Bernoulli variable. The model assumes that user interest in a KC (topic) is a probability where values are close to 1, which indicates a high probability of interest. The model aims to predict the engagement e_{ℓ_I, r_x}^t of the learner ℓ with resource r_x at time t , which is based on the learner's interest $\theta_{\ell_I}^t$ in KCs K_{r_x} and their presence i_{r_x} in the resource as:

$$\hat{e}_{\ell, r_x}^t = \begin{cases} 1 & \text{if } P(e_{\ell_I, r_x}^t | \theta_{\ell_I}^t, K_{r_x}, i_{r_x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The graphical model representing the interest tracing model is presented in Figure 4 (left) and is an adaptation of an online knowledge tracing model (online KT model) [10] that is inspired by an online skill assessment model [52]. As per Figure 4 (left), we compute $P(e_{\ell_I, r_x}^t | \theta_{\ell_I}^t, K_{r_x}, i_{r_x})$ by computing the joint probability of learner ℓ 's interest in KCs K_{r_x} as $\text{AND}(\theta_{\ell_I}^t)$, $j \in K_{r_x}$ together with a noise factor. The noise factor accounts for the two scenarios where (i) the learner has interest and decides to disengage due to other reasons (e.g., lack of time, contextual limitations, etc.), or when (ii) the learner lacks interest but decides to engage (e.g., an academic obligation). Similar factors have been taken into account to address the guess and miss situations in question by answering/knowledge tracing scenarios [30,52]. As the parameters are Bernoulli and the outcome is binary, we use loopy belief propagation [68] to execute inference steps in this model.

3.4.2. TrueLearn Interest Model

TrueLearn Interest is a reformulation of the IRT-inspired TrueLearn fixed-depth model [10], which is furthermore inspired by the TrueSkill model [51]. The model tries to determine if the learner ℓ has higher interest in topics K_{r_x} that are present in the resource r_x (binary i_{r_x}) as per Figure 3i. Figure 4 (middle) illustrates the factor graph of the TrueLearn Interest model. In contrast to the interest tracing model, TrueLearn Interest uses Gaussian variables where the level of interest (μ_{ℓ}) and degree of uncertainty (σ_{ℓ}^2) are inferred. The i_{r_x} are binary values that are observable ($i_{r_x} \sim \mathcal{N}(1, 0^2)$). The β^2 factor acts as the noise factor. The prediction \hat{e}_{ℓ, r_x}^t is made similarly using the criteria outlined in Equation (3). As per Figure 4 (middle), the prediction probability is calculated using:

$$P(e_{\ell_I, r_x}^t | \theta_{\ell_I}^t, K_{r_x}, i_{r_x}) = \text{CDF}_{\mathcal{N}}(p_{\ell_I, r_x}^t - p_{r_x} \geq 0.0). \quad (4)$$

In other words, it is validated if the learner interest $\theta_{\ell_I}^t$ in topics K_{r_x} is significantly larger than the sum of their presence i_{r_x} values. This result is a normal variable because the result is a difference between two Gaussian variables. The cumulative distribution function ($\text{CDF}_{\mathcal{N}}$) can be used to estimate the probability that the learner interest is larger than zero.

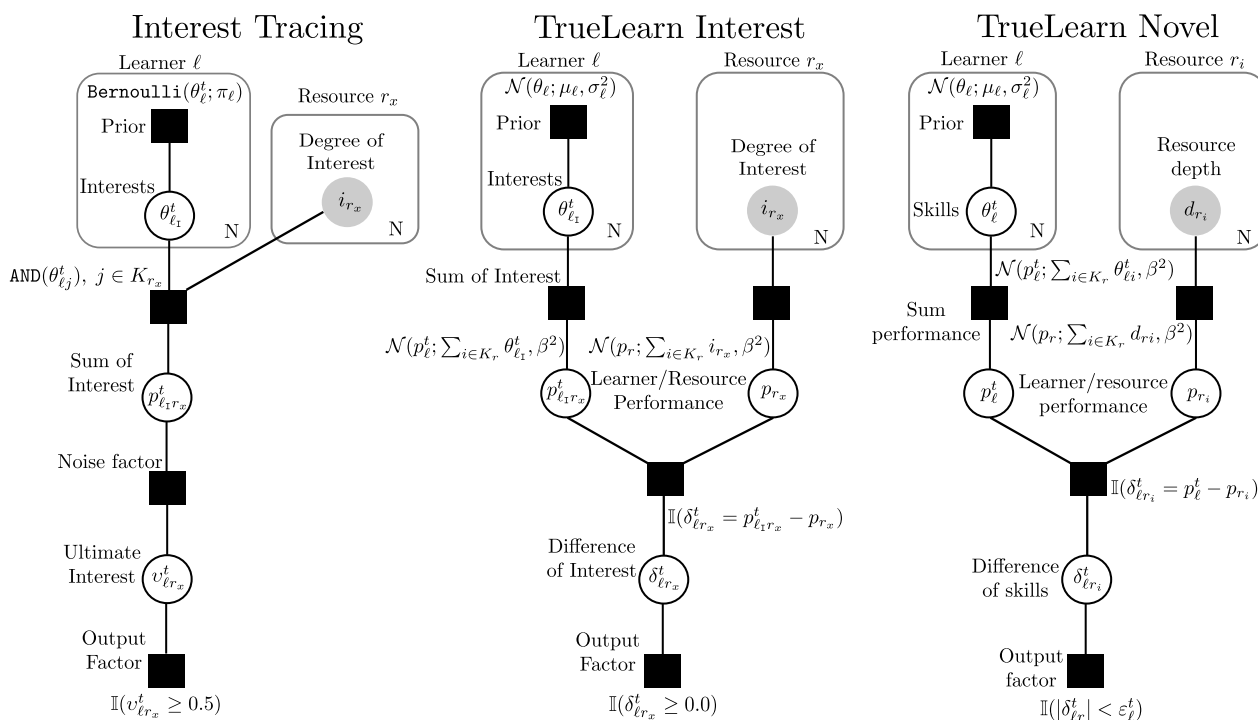


Figure 4. Factor graphs representing the probabilistic graphical models for interest tracing (left), TrueLearn Interest (middle), and TrueLearn Novel (right) models.

3.5. Combining Interest, Novelty, and Knowledge: TrueLearn INK Model

We use the TrueLearn Novel model to infer the knowledge representation of the learner while modelling novelty. It is one of the most recent learner models that accounts for both the novelty of content and the knowledge state of the learner while demonstrating superior performance when predicting engagement with video lectures [10,40]. The TrueLearn Novel model is an online model using density filtering and recovers learner skills even if the concepts occur sparsely [10]. This makes the algorithm scalable and data efficient. The model also uses entity linking for content annotation by creating an interpretable user model. The hypothesis used by the TrueLearn Novel models is illustrated in Figure 3ii, while the factor graph of the TrueLearn Novel model is presented in Figure 4 (right). Being similar to the TrueLearn Interest model described earlier, the TrueLearn Novel model also uses expectation propagation [69] for inference. We direct the reader to the original TrueLearn Novel manuscript for more details [10].

Given that we have two distinct graphical models that predict learner engagement based on two independent variables: (i) interest (e_{ℓ_I, r_x}^t) and (ii) novelty and knowledge (e_{ℓ_{NK}, r_x}^t), we experiment with combining the predictions to create the TrueLearn INK model as per Equation (5).

$$e_{\ell_{INK}, r_x}^t = f(e_{\ell_I, r_x}^t, e_{\ell_{NK}, r_x}^t) \tag{5}$$

where f is a function that combines the predictions. To choose $f(\cdot)$, we experimented with multiple ensemble-learning techniques belonging to two main groups.

- Probabilistic Combination of Outcomes: Using probability theory to combine the predictions together;
- Meta-Learner: Learning how to weigh the two predictions to obtain a more accurate final engagement prediction.

3.5.1. Using Probabilistic Combination with Existing Meta-Learners

In the context of probabilistic combination, we tried both the (i) AND($e_{\ell_I, r_x}^t, e_{\ell_{NK}, r_x}^t$) operator, the more restrictive assumption where both models have to predict high probabilities for the ultimate prediction e_{ℓ_{INK}, r_x}^t to be positive, and (ii), the OR($e_{\ell_I, r_x}^t, e_{\ell_{NK}, r_x}^t$) operator, the less restrictive assumption [60].

When choosing a meta-learner, we restricted ourselves to online learning schemes to preserve the scalability of TrueLearn INK. We also dismissed multi-layer perceptrons for maintaining a linear meta-learner that increased the interpretability of the learned weights. Perceptron and stochastic logistic regression models are selected for experimentation.

3.5.2. Meta-TrueLearn

Additionally, we formulated a linear model that learned weights using an expectation propagation analogue compared with the TrueLearn models. The meta-TrueLearn model can be seen as a Bayesian probit regressor. A similar weight-learning scheme has previously been used in a personalised click-through prediction task [70]. The proposed meta-learning model is illustrated in Figure 5. As seen in the figure (only black parts), two predictions from the interest $e_{\ell_I}^t$ and TrueLearn Novelty model $e_{\ell_{NK}}^t$ are fed as observed constants (i.e., Gaussians with zero variance). The product factor and sum factor are used to model the final engagement $e_{\ell_{INK}}^t$ as:

$$e_{\ell_{INK}, r_x}^t = \mathbf{W}_{\ell_I}^t \cdot e_{\ell_I, r_x}^t + \mathbf{W}_{\ell_{NK}}^t \cdot e_{\ell_{NK}, r_x}^t + \mathbf{b}_{\ell}^t \tag{6}$$

where $\mathbf{W}_{\ell_I}^t$, $\mathbf{W}_{\ell_{NK}}^t$, and \mathbf{b}_{ℓ}^t are trainable parameters that are modelled as Gaussian variables, while \mathbf{b}_{ℓ}^t is the bias term. The model trains in a greedy fashion, contributing to computational efficiency. Although the prediction step is executed in every time step, weight updates only happen when a misclassification is encountered.

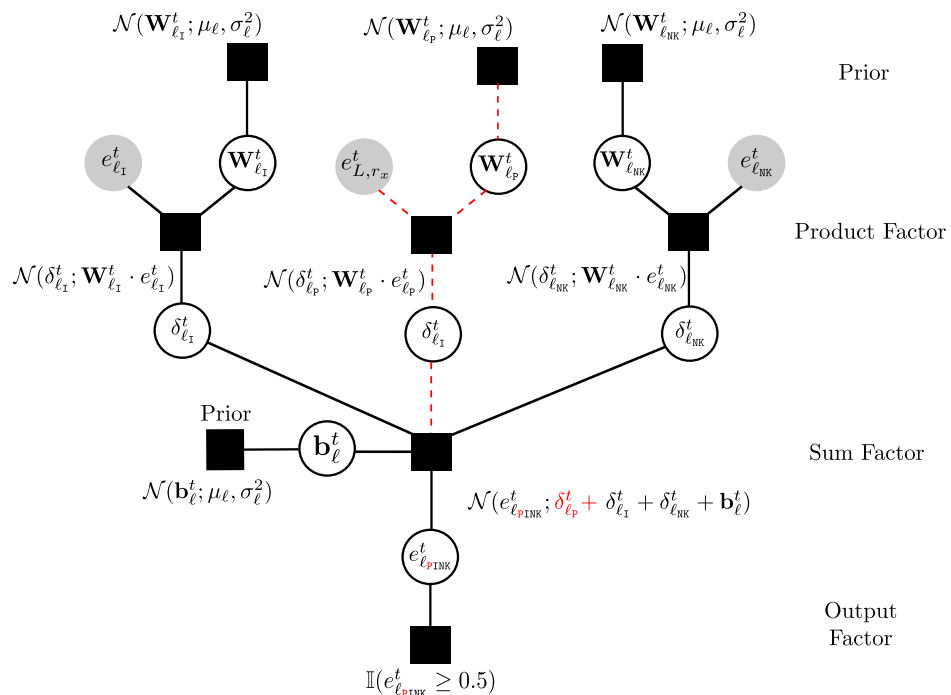


Figure 5. Factor graphs representing meta-TrueLearn, a probabilistic graphical model to combine the predictions from both interest and knowledge-based engagement models. Optionally, the same meta-learner can be used to combine the population-based model (parts highlighted in red dashes).

3.6. Combining Population-Based Prior (P + INK): TrueLearn PINK Model

One of the weaknesses of an online learning scheme such as TrueLearn is that it starts learning the model using historical interaction data from the user. This means that these models do not have any historical data at the beginning (when t is close to 1, hereafter referred to as t_{small}), leading them to use non-informative priors (Gaussians with mean close to 0 and a large variance) when calculating $e_{\ell_I}^{t_{\text{small}}}$ and $e_{\ell_{\text{NK}}}^{t_{\text{small}}}$. This causes TrueLearn INK to estimate the same probability of engagement $e_{\ell_{\text{INK}}, r_x}^{t_{\text{small}}}$ for every user regardless of what video fragment r_x they watch in the first interactions with the system, implying a cold-start problem. Popularity-based (context-agnostic) recommenders are one remedy to the cold-start problem [58].

Several of our recent works [3] have proposed methods leveraging content features and topical features (as per Figure 1) to predict population-based engagement of an educational video. We use this approach to build a context-agnostic engagement predictor for video lectures, estimating e_{L, r_x} , engagement of the learner population L with resource r_x , creating a prior for engagement where learner knowledge and interest-based estimates are non-informative. We hypothesise that using such an informative prior in the early phase of a user session and gradually transitioning to rely on learned personalised interest and knowledge models (TrueLearn INK) can address the cold-start problem. We define our final proposal TrueLearn PINK, a hybrid recommender model that starts engagement prediction heavily relying on a population-based prior and transitions to TrueLearn INK over time. We devise two approaches used in hybrid recommendation systems [64], (i) switching and (ii) stacking (i.e., meta-learner) to facilitate the transition from population prior to TrueLearn INK.

3.6.1. TrueLearn PINK (Switching)

This algorithm starts making engagement predictions using the population-based engagement predictor and switches after n events. Algorithm 1 is used for each learner $\ell \in L$ in TrueLearn PINK (switching).

Algorithm 1 Hybrid Recommender TrueLearn PINK using Switching

Require: $0 \leq e_{\ell_I}^t, e_{\ell_{\text{NK}}}^t \leq 1$

Require: $n \geq 1$ ▷ upper ceiling of t_{small}

Ensure: $t \geq 1$

for $t \in \{1 \dots T_\ell\}$ **do**

if $t \leq n$ **then** ▷ t_{small} scenario

$e_{\ell_{\text{PINK}}}^t \leftarrow e_{L, r_x}$ ▷ estimate from population-based predictor

else if $t > n$ **then**

$e_{\ell_{\text{PINK}}}^t \leftarrow e_{\ell_{\text{INK}}, r_x}^t$ ▷ estimate from personalised model

end if

end for

3.6.2. TrueLearn PINK (Meta)

In contrast, this approach maintains a set of weights that determine how much influence the different predictions (from population-based to interest and knowledge models) have on the final prediction. This is performed by using the same meta-learner described in Section 3.5. We included a new additive component $\mathbf{W}_{\ell_p}^t \cdot e_{L, r_x}$ in Equation (6) to modify the factor graph (including red-dashed components of Figure 5). $\mathbf{W}_{\ell_p}^t$ is a trainable parameter.

3.7. Experiments

We ran a series of experiments that allowed us to answer multiple research questions related to the performance of the proposed TrueLearn models.

- RQ 1: How well do the interest models perform?
- RQ 2: How well do different combining mechanisms perform with TrueLearn INK?
- RQ 3: Does combining the individual models lead to superior performance?
- RQ 4: Does combining the population-based component in early stage prediction further improve performance?

We designed our experiments in a phased methodology. The sequential integration of different factors is graphically illustrated in Figure 6. The systematic integration of TrueLearn Interest, TrueLearn Novel, and the population-based model leads to the final TrueLearn PINK model. As per the figure, the experiments around our contributing models helped us to resolve the relevant research questions.

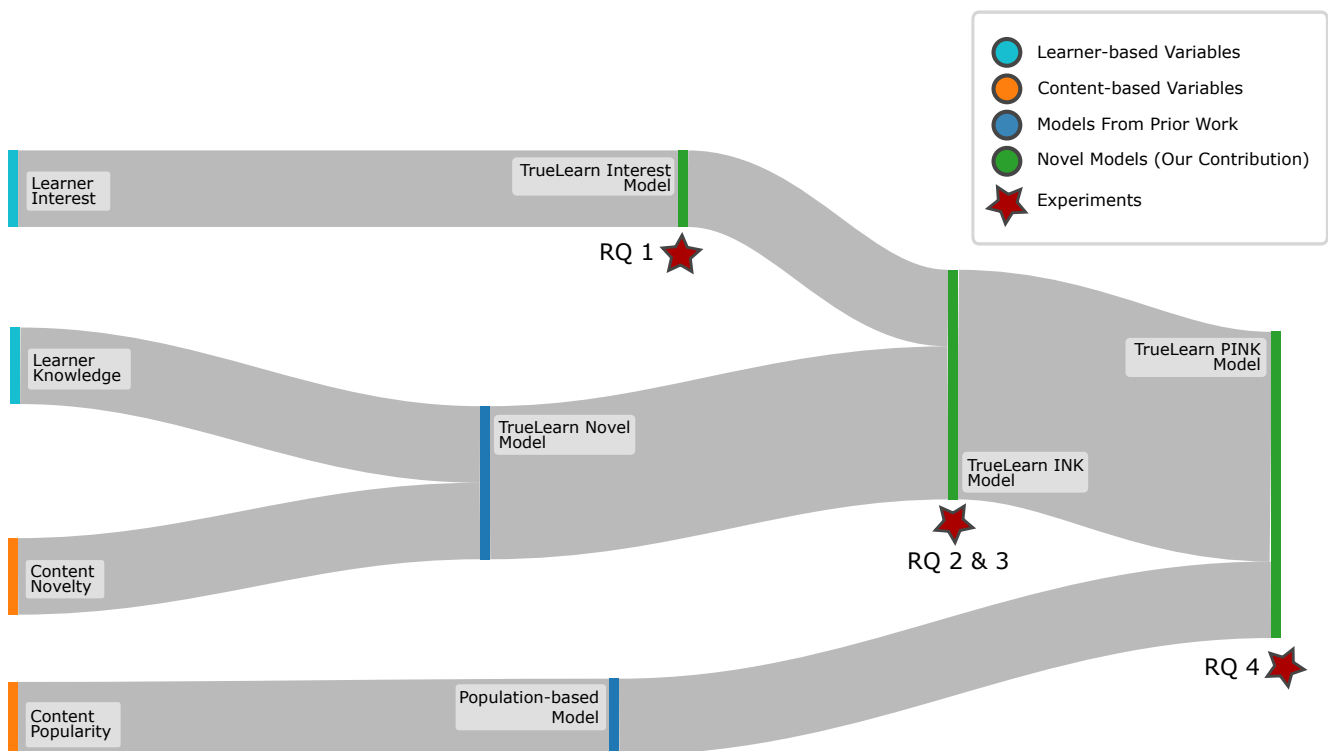


Figure 6. Graphical illustration of the experimental design where we sequentially integrated (I)nterests, (N)ovelty, (K)nowledge, and (P)opularity factors together, creating a unified model.

In the first experiment addressing RQ 1, the two proposed models, namely (i) interest tracing and (ii) TrueLearn Interest, are compared with the six baseline interest models described in Section 3.3, and the results are reported in Table 1.

In the second experiment addressing RQ 2, we took the best-performing interest model from the former experiment and combined its predictions with the knowledge and novelty model, creating TrueLearn INK (results in Table 2). In terms of accounting for novelty and knowledge, the TrueLearn Novel [10] is used where the number of KCs is kept to $|K_{r_x}| = 3$, which is based on the results demonstrated with the PEEK dataset [40]. To identify the best ensemble approach for TrueLearn INK, we benchmarked multiple methods. From the probabilistic approach, we tried both the AND and OR assumptions. We empirically evaluated three online meta-learners, namely (i) Logistic, (ii) Perceptron, and (iii) Meta-TrueLearn. When experimenting with meta-learners, we experimented both with and without the bias term to identify the best model. As a post-step, we further compared the predictive performance of isolated models vs. the combined model to identify the strengths and weaknesses of the TrueLearn INK model, thereby addressing RQ 3. The results of the post-analysis are found in Figure 7.

Finally, we launched our third experiment once the best combiner for the TrueLearn INK model had been determined. We further combined the population-based engagement prediction model with it. This combination is experimented on by switching and stacking (using a meta-learner). We call this model TrueLearn PINK. The results of TrueLearn PINK are reported in Table 3.

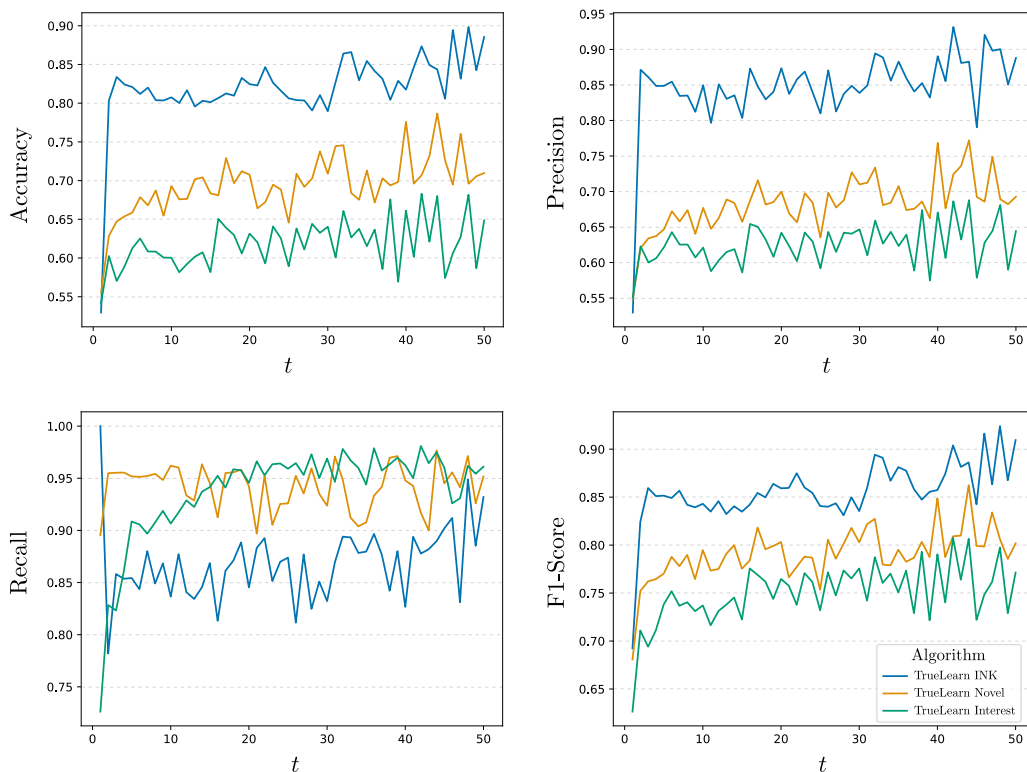


Figure 7. How the mean accuracy, precision, recall, and F1 score at time t across all users change on TrueLearn Interest (green), TrueLearn Novel (yellow), and TrueLearn INK Meta (blue) models.

Table 1. Weighted average test-set performance for accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 score (F1). The best-performing and next best values are highlighted in **bold** and *italic* faces, respectively. The proposed models that outperform baseline counterparts in the PEEK dataset ($p < 0.01$ in a one-tailed paired t -test) are marked with $\cdot(*)$.

Algorithm		Acc.	Prec.	Rec.	F1
Baseline Models	Cosine	55.08	57.86	58.45	54.06
	Jaccard _C	55.46	57.81	60.36	55.03
	Jaccard _U	64.05	<i>57.85</i>	72.76	<i>61.22</i>
	TF(Binary)	55.19	56.71	66.60	57.38
	TF(Cosine)	55.11	56.75	65.95	57.11
	TFIDF(Cosine)	41.80	31.70	9.05	10.67
Our New Proposals	Interest Tracing	47.95	52.05	37.24	38.96
	TrueLearn Interest	<i>57.70</i>	56.83	78.74 (*)	62.50 (*)

Table 2. Weighted average of PEEK dataset test-set performance for accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 score (F1). The best-performing and next best values are highlighted in **bold** and *italic*, respectively. The proposed models that outperform baseline counterparts in the PEEK dataset ($p < 0.01$ in a one-tailed paired t -test) are marked with $\cdot^{(*)}$.

Algorithm	Acc.	Prec.	Rec.	F1
Best Baselines from Table 1				
TF(Binary)	55.19	56.71	66.60	57.38
Jaccard _U	64.05	57.85	72.76	61.22
TrueLearn Models in Isolation				
TrueLearn Interest	57.70	56.83	78.74	62.50
TrueLearn Novel	64.40	58.42	80.15	65.12
TrueLearn INK Models (Our New Proposals)				
AND	65.33 $\cdot^{(*)}$	58.70 $\cdot^{(*)}$	69.80	61.68
OR	56.74	56.74	88.92 $\cdot^{(*)}$	65.63 $\cdot^{(*)}$
Logistic	78.58 $\cdot^{(*)}$	64.07 $\cdot^{(*)}$	68.17	65.86 $\cdot^{(*)}$
Perceptron	78.56 $\cdot^{(*)}$	64.05 $\cdot^{(*)}$	68.58	66.04 $\cdot^{(*)}$
Meta-TrueLearn	78.71 $\cdot^{(*)}$	64.19 $\cdot^{(*)}$	68.62	66.14 $\cdot^{(*)}$

Table 3. Weighted average of PEEK dataset test-set performance for accuracy (Acc.), precision (Prec.), recall (Rec.), and F1 score (F1). The best-performing and next best values are highlighted in **bold** and *italics*, respectively. The proposed models that outperform baseline counterparts in the PEEK dataset ($p < 0.01$ in a one-tailed paired t -test) are marked with $\cdot^{(*)}$.

Algorithm	Predicting First Event				Predicting All Events			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Best Performing Model from Table 2								
TrueLearn INK	44.21	44.21	100.0	61.32	76.26	63.36	69.30	65.84
TrueLearn PINK Models (Our New Proposals)								
Switching	56.09 $\cdot^{(*)}$	50.32 $\cdot^{(*)}$	53.58	51.89	77.08 $\cdot^{(*)}$	63.92 $\cdot^{(*)}$	66.55	64.95
Meta	56.02 $\cdot^{(*)}$	50.25 $\cdot^{(*)}$	53.58	51.85	78.90 $\cdot^{(*)}$	64.88 $\cdot^{(*)}$	66.06	65.29

3.8. Evaluation Metrics

We used a hold-out validation technique in all the experiments where the training data is used for hyperparameter tuning. We used the test data to measure and report the predictive performance. As the downstream prediction task predicts discrete labels [40], we used classification metrics, namely accuracy, precision, recall, and F1 score, to report the performance. F1 score and accuracy are used as the primary metrics for model comparison as they represent correctly predicting both positive and negative labels. When evaluating with the full PEEK dataset, a learner-wise paired t -test is used to measure if the proposed models lead to a statistically significant improvement over the baselines. Pearson R correlation is used where correlation analysis between two continuous variables is performed.

The results reported in Tables 1 and 2 use learner events $2 \leq t \leq T_\ell$ for computing evaluation metrics. This is because all the baselines and our proposals in these tables are not capable of making meaningful predictions when historic data is absent (when $t = 1$). However, TrueLearn PINK addresses this weakness. Hence, the results in Table 3 use learner events $1 \leq t \leq T_\ell$ for computing evaluation metrics.

4. Results and Discussion

The results from all the different experiments outlined in Section 3.7 are presented in this section. Our experiments focus on understanding the prediction quality of the proposed models, namely TrueLearn Interest, TrueLearn INK, and TrueLearn PINK (refer to Figure 6). All TrueLearn models are massively parallelisable because the learner model

for different learners can be learned independently. Our current implementation is in Apache Spark [71] (using the MapReduce paradigm [72]), which demonstrates this fact. All experiments were run with an AMD Ryzen Threadripper 2950X CPU with 15GB memory. The code to reproduce the proposed models and experiments is available online (<https://github.com/sahanbull/TrueLearn>, accessed on 1 May 2022).

4.1. Predictive Performance of TrueLearn Interest (RQ 1)

Different numbers of topics $|K_{r_x}|$ are experimented with to identify the best performing $|K_{r_x}|$ for the interest tracing and TrueLearn Interest models. Empirical results showed that the most predictive feature settings were $|K_{r_x}| = 1$ and $|K_{r_x}| = 5$ for interest tracing and TrueLearn Interest, respectively. Table 1 outlines the predictive performance of the proposed interest models in comparison with the relevant baselines. The table shows clear evidence of the superiority of the TrueLearn Interest model in comparison with the baseline interest models with the PEEK dataset. The results show that the performance of TrueLearn Interest is significantly better than the next best performing baseline in terms of recall and F1 score. This model also surpasses all but one of the baselines when it comes to accuracy. The performance of the proposed interest tracing model is not competitive. This difference of performance observed between the Bernoulli- and Gaussian-based models is consistent with the results obtained in our prior work [10,40] when evaluating similar models for modelling knowledge.

4.1.1. On Performance of Jaccard_U Model

The results in Table 1 show that the Jaccard_U model contends closely with the TrueLearn Interest model. However, the TrueLearn Interest model significantly outperforms Jaccard_U in the full PEEK dataset in terms of F1 score, which combines precision and recall. On the other hand, the Jaccard_U model is much better at predicting negative engagement, as shown by the superiority of the accuracy score. From a data efficiency perspective, Jaccard_U has several weaknesses. As the Jaccard score is computed based on the number of users who visited videos, the Jaccard_U model really struggles when the number of users is small [40]. While individual users could be very active, the Jaccard_U model will not work unless many of them co-visit videos. In contrast, the TrueLearn Interest model does not have this issue as it does not rely on other users' actions for recommendation or content representation.

4.1.2. TrueLearn Interest vs. TrueLearn Novel

The experimental results demonstrate that TrueLearn Interest (as per Table 1) and TrueLearn Novel [40] are the most suitable choices when accounting for the interest, novelty, and knowledge of learners in predicting engagement. However, the results in Table 2 also show that the performance of these two models are very similar to each other. Therefore, it is sensible to validate if the prediction behaviours of these two models are significantly different from each other. To validate this, we take the predictions made on the test set by both the models separately and measure the agreement between their engagement predictions as:

$$\text{Agreement} = \frac{1}{\sum_{\ell \in L} T_{\ell}} \sum_{\ell \in L} \sum_{t=1}^{T_{\ell}} \mathcal{A}(e_{\ell_{I,r_x}}^t e_{\ell_{NK,r_x}}^t), \quad (7)$$

where $\mathcal{A} = \begin{cases} 1 & \text{if } e_{\ell_{I,r_x}}^t = e_{\ell_{NK,r_x}}^t \\ 0 & \text{otherwise} \end{cases}$, T_{ℓ} is the number of events of learner ℓ , and L is the full set of learners.

The experimental results show that there is only 73.1% agreement between the predictions coming from the two models. Therefore, we can observe that there is a significant deviation between the behaviours of the two models. This observation further reinforces the utility of combining the two hypotheses together.

4.2. Predictive Performance of TrueLearn INK (RQ 2 and 3)

The predictive performances of the different versions of TrueLearn INK using different methods of combination are outlined in Table 2. This table also includes the two best-performing baselines from Table 1, in addition to the performances of the two TrueLearn models when they are used in isolation. The primary observation from Table 2 is that the majority of the combination approaches (except AND model) lead to significantly improved accuracy, precision, and F1 score. Among the TrueLearn INK models, the meta-learning approaches seem to perform better than the probabilistic combination approaches.

Among the probabilistic approaches, the AND model tends to be very restrictive when predicting positives, leading to improved accuracy and precision. On the other hand, the OR model seems to be too relaxed, leading to a significantly higher recall at the cost of degraded accuracy and precision.

Table 2 further shows that the meta-learner-based approaches seem to be equally competent in prediction, with minor differences in performance. Our proposal, the meta-TrueLearn model, is the best-performing model among its meta-learning counterparts. The meta-TrueLearn-based TrueLearn INK model gives the best performance among all the tested models in terms of accuracy, precision, and F1 score, while attaining the highest recall score among the meta-learner-based models. Table 2 also shows that the meta-learning models lead to more accurate and precise models, pushing recall lower in comparison with using the TrueLearn Interest and TrueLearn Novel models in isolation. This observation is further confirmed by the performance reported in Figure 7.

We take the prediction for the t^{th} event of the entire test-set population and calculate the performance metrics for the three models to create Figure 7. We restrict the plots to $t \leq 50$ because the number of learners with more events is very small [40], making the metrics unstable. The figure shows how TrueLearn models rapidly learn to predict with accuracy, precision, and F1 boosting from ≈ 0.5 to 0.85 in the first five events. This figure further elaborates how the prediction quality of TrueLearn INK is superior throughout the learner journey (both early and later stages) in comparison with using the models in isolation. The only exception is recall, where the TrueLearn INK model is inferior. This is expected as the weight updating mechanism of meta-TrueLearn model is driven by accuracy.

Meta-Weights and Topic Sparsity

In the deeper analysis, we observed that there is a relationship between how the meta-weights $\mathbf{W}_{\ell_1}^t$ and $\mathbf{W}_{\ell_{\text{NK}}}^t$ change with the number of unique topics/KCs a learner encounters. We plotted Figure 8 by taking the final weights of the user $\mathbf{W}_{\ell_1}^T$ and $\mathbf{W}_{\ell_{\text{NK}}}^T$ and plotting it against the total number of unique KCs for the user. Pearson R correlation analysis between these variables also showed that the number of unique topics has a positive correlation of 0.028 ($p \leq 0.05$) and a negative correlation of -0.234 ($p \leq 0.01$) with $\mathbf{W}_{\ell_1}^T$ and $\mathbf{W}_{\ell_{\text{NK}}}^T$, respectively. This suggests that the model learns to concentrate on an interest more when a large number of new topics are encountered.

4.3. TrueLearn PINK: Addressing the Cold-Start Issue for TrueLearn INK (RQ 4)

In Figure 9 (depicted in dark blue), the average performance of TrueLearn INK on the first 10 events of the learner test-set sample clearly shows that it struggles in the first few (≈ 5) events of each user. This is an expected observation because TrueLearn INK uses no information to estimate the engagement of the user at time step 1. This cold-start problem needs to be addressed in order to improve the effectiveness of personalisation.

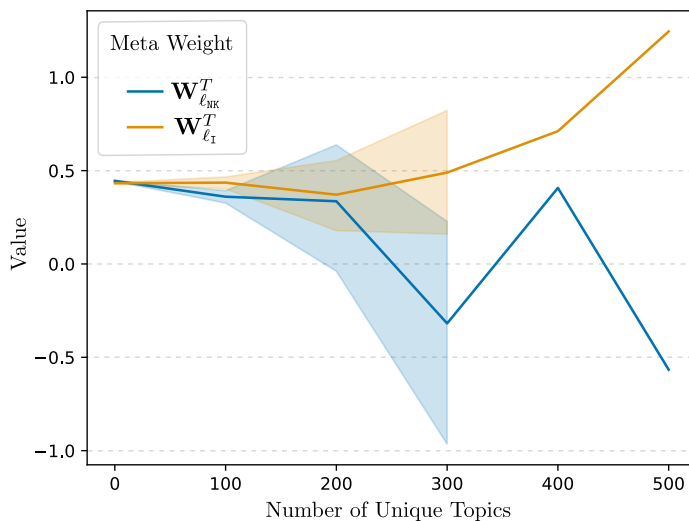


Figure 8. Changes in meta-weights $W_{l_I}^T$ (Orange) and $W_{l_{NK}}^T$ (Blue) with respect to the number of unique topics for each learner l in the test set

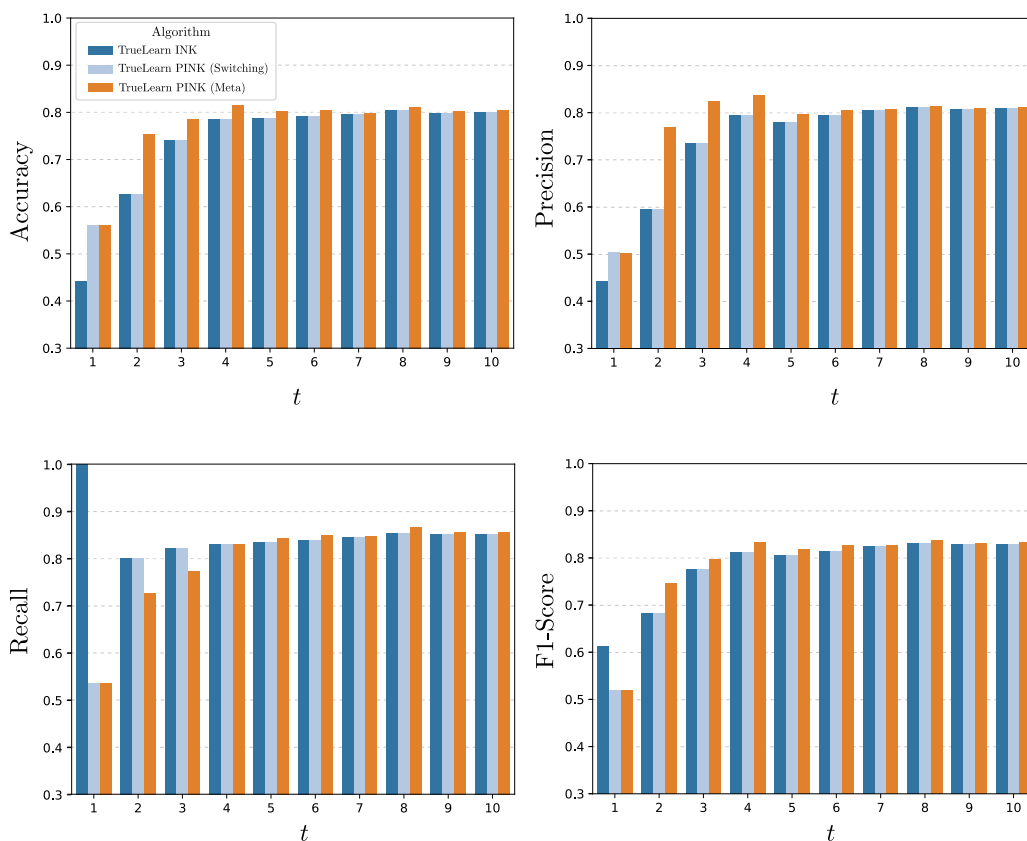


Figure 9. How accuracy, precision, recall, and F1 score of TrueLearn INK (dark blue); TrueLearn PINK (Switching) (light blue), which uses switching approach; and TrueLearn PINK (Meta) (yellow), which uses a meta-learner approach, changes over time step t across the entire learner test-set population.

We address the cold-start weakness of TrueLearn INK described in Section 4.2 by combining a population-based prior, creating TrueLearn PINK. Table 3 shows the performance of TrueLearn PINK when combining the population-based prior using (i) switching, as per Algorithm 1, and (ii) a meta-learner, as per Figure 5. It is evident from Table 3 that this

enhancement leads to significant improvements of accuracy and precision while sacrificing recall, which also leads to a drop in the F1 score.

However, TrueLearn PINK behaves differently in the first event because the model has additional prior information (coming from the population-based prediction model). Both the accuracy and precision of the predictions in the first event of the learner population significantly improves (in Table 3, left part); furthermore, the recall will fall, as the proposed context-agnostic model only captures a population-based prior, which may deviate from the individuality of the learners. However, it can be argued that making a prediction with partial additional information is still better than predicting with no prior information. In the bigger picture (in Table 3, right part), being able to make slightly more informed and varied predictions for the early events of the learners based on lecture content features enables significant improvements in the prediction accuracy and precision of TrueLearn PINK on the entire test set.

A magnified look at the performance of the first 10 events of the learners outlined in Figure 9 further elaborates the effect of TrueLearn PINK. TrueLearn PINK (Switching) in light blue, only using the population prior in the first event ($t_{small} \in 1$ or $n = 1$ determined via grid search) shows superior accuracy and precision to TrueLearn INK, and then matches its performance as the algorithm switches from $t \geq 2$. The TrueLearn PINK (Meta) in orange maintains a learnable set of meta-weights for popularity, interest, and knowledge models (set to $\mathbf{W}_{\ell_p}^{t=1} = 0.90$, $\mathbf{W}_{\ell_I}^{t=1} = 0.05$, and $\mathbf{W}_{\ell_{NK}}^{t=1} = 0.05$ in the beginning). As the population-based model does not get suppressed in event 2 when using TrueLearn PINK (Meta) (in contrast to TrueLearn PINK (Switching)), the superiority of accuracy and precision persists in the early stage (t_{small} scenario) while also beating its counterparts in F1 score when $t \geq 2$.

We encounter further interesting observations from Figure 9 and Table 3. Firstly, the TrueLearn PINK (Meta) model leads in predictive performance when it comes to predicting learner engagement across events in the entire dataset. Explaining this observation with Figure 9 is fairly straightforward as the meta-learner approach involves the population-based prior in multiple events that go beyond the first event of each learner. This indicates to us that using the population-based prior in multiple events followed by the first event is more beneficial as the TrueLearn models cannot learn reliable interest and knowledge parameters by using only one interaction. However, it gets interesting as the same meta-learner model is outperformed by the TrueLearn PINK (switching) model in the first event (refer to Table 3). This means that giving 100% weighting to the population-based model's prediction in the switching approach rather than 90% in the meta-learner approach leads to significantly better predictions in the first event. This observation is also explainable. In the meta approach, we fuse the population-based predictions with two non-informative predictions (90:5:5 in the beginning), where the non-informative predictions add noise to the final prediction. However, adding 5% weight to these non-informative predictions is essential to learn the weighting of the contextual predictions that are important for personalisation. This observation further proves that the population-based model gives a refined signal that the model can use in the early events.

Impact of the Population-Based Model

By observing the left part of Table 3, one can determine that TrueLearn INK always predicts positive engagement for each learner at time step 1. At the first event, the recall of TrueLearn being 1.0 while the accuracy and precision are the same depicts this fact. TrueLearn predicts positively in each user's first event because the model has no information to base the prediction on and, hence, uses the initial mean and variance (set as hyperparameters) to make the first estimate [10]. This will output the same probability value for any learner ℓ at time $t = 1$, regardless of what video fragment r_x they are about to engage with. As these models are intended to be used for recommendations, having a $t = 1$ prediction that gives a uniform score for every material takes the ranking ability away from the model. In other words, the TrueLearn INK model will rank the resources randomly to each learner in the first event. Therefore, having the capability to estimate

different scores to different materials in the first step will definitely help a recommender, as it allows ranking materials. Table 3 and Figure 9 further confirm that TrueLearn PINK is capable of scoring different video fragments r_x differently when $t \in t_{\text{small}}$, and is capable of practically ranking materials in the system while gaining performance.

4.4. Opportunities and Limitations

As Section 4 suggests, TrueLearn INK and TrueLearn PINK bring significant performance gains over the baseline recommendation systems by combining the content popularity, learner interests, and novelty and knowledge factors together (as per Figure 1). The proposed Bayesian models learn rapidly with a very small number of events, making them data efficient. They also exclusively rely on the individual learners' data, making them massively parallelisable and privacy preserving by design. In a new age where OERs can be resourced for personalised education at a global scale, the data efficiency, privacy, and scalability features that the TrueLearn family of models possess make it a strong candidate for ensuring quality education while respecting the individual liberties and rights of global citizens. Its reliance on Wikification [32] for creating content representations also allows scalable content annotation, which is pivotal to the process of rapidly including large numbers of newly created materials to the OER collection, keeping it up-to-date. Wikipedia, which is multi-lingual, cross-domain (from sciences to arts), and temporally-dynamic (updating its knowledge with time), enables the building of content representation that is robust to the evolution of universal knowledge itself [1].

The human-intuitive learner representation that all TrueLearn models maintain is also something that is worthy of attention. Having a representation that relies on human-intuitive symbols (Wikipedia topics in this case) allows the AI-powered system to communicate with the human user with actionable narratives [73]. Such human-intuitive learner models have proven to be key to promoting self-reflection and meta-cognition in learners [74]. Such expressive models also pave a pathway to facilitating active thinking and self-regulated learning [75], which are critical triggers for success in an informal, lifelong e-learning scenario. Human intuitiveness goes far beyond presenting insights to the learner. The presentation can be further extended to a two-way communication channel between the human learner and the AI system. Users could theoretically check the latent variables of TrueLearn models to understand what the model believes about them, encouraging the users to intervene and change the model's perception (e.g., by correcting and repositioning skill/interest parameters for different KCs). The same is applicable to controlling the ensemble of hypotheses that formulate the final prediction, as it gives the user agency to decide what drives their recommender system. For instance, the learner could decide to weight their interest more than the knowledge state, depending on their immediate needs. In summary, the transparent representations allow for the development of different levels of human-in-the-loop AI within the recommender [76]. In the eyes of policy makers and scientists, such transparency can serve as a powerful diagnostic tool that will allow for the scrutinisation of the model's assumptions and incorporating other non-technical features, such as explainability, accountability, and fairness, which are essential for a global-scale educational recommender that can shape the future of the broader world population.

While TrueLearn is developing in a promising direction, it has limitations that demand attention. While TrueLearn proves excellent in predicting the engagement of learners, engaging with learning materials does not explicitly translate into users acquiring new knowledge or attaining learning gains, which is the key objective of learning. While learner engagement correlates with learning gains [77], explicit question answering is the obvious way to verify knowledge; however, it is not a feature of the TrueLearn model at this time. The immaturity of automatic question generation models partly contributes to this limitation. In addition, current TrueLearn models assume that the KCs they model are independent from each other, despite Wikipedia-based KCs being semantically related [78]. This is a major weakness of the model assumptions that needs to be addressed in future work. While video-watching time is used by TrueLearn, many other interaction patterns

(e.g., pausing, replaying, and skipping ahead) that carry stronger signals of engagement are not used. TrueLearn's ability to work with non-video OERs (PDFs and interactive materials) is also not developed.

5. Conclusions

This work marks a step towards creating integrative educational recommendation systems that can account for different aspects when predicting learner engagement by modelling the user state. We hypothesise that this can not only improve predictive performance by considering the multiple factors involved in engagement but could also help improve the explainability of recommendations. We make multiple model contributions: (i) TrueLearn Interest, a competitive learner interest representation; (ii) TrueLearn INK, an integrative learner model that accounts for learner interest, and novelty and knowledge; and (iii) TrueLearn PINK, which uses a population-based prior to address the cold-start problem in educational recommendation processes. Our extensive experiments show that the proposed models significantly improve on the competitive baselines while retaining the important qualities of a lifelong-learning educational recommender, such as scalability, data efficiency, and transparency. Experiments with interest models suggest that the Gaussian-based TrueLearn Interest model is most suitable. Experiments with combining two independent TrueLearn models also lead to a significant improvement of predictive performance over using the models in isolation. We also observe that meta-learner-based ensemble approaches tend to perform better among different combining techniques. The proposed meta-TrueLearn, an online Bayesian probit regression scheme, exhibits the best predictive performance among the TrueLearn INK and PINK models. The analysis of this model shows that the ensemble model improves overall performance by significantly improving accuracy and precision while sacrificing some recall. Further experiments in combining the population-based engagement predictor as a prior for cold-start scenarios show that the combined TrueLearn PINK model can push the accuracy and precision of the model further than TrueLearn INK.

Empowering educational recommendation technology still has much scope for further development. Running extensive experiments on how to incorporate semantic relatedness between KCs is our immediate priority. Additional model assumptions, such as interest dynamics and forgetting [79], also need to be explored and accounted for. Another major missing piece that will truly boost the utility of the TrueLearn recommender is developing how the model can interface with human users to become a truly collaborative/cooperative AI rather than a prescriptive one. Finally, implementing the recommender to drive OER-based learning platforms, such as X5Learn [21], will allow us to run larger scale, in-the-wild user studies while also focusing on much needed extensions, such as working with non-video materials and scalable question generation for personalised testing. We strongly believe that TrueLearn, when further enriched as discussed above, could help deliver equitable personalised online education.

Author Contributions: Conceptualization, S.B. and M.P.-O.; Data curation, S.B.; Formal analysis, S.B.; Funding acquisition, E.Y. and J.S.-T.; Investigation, S.B. and M.P.-O.; Methodology, S.B.; Project administration, E.Y. and J.S.-T.; Resources, E.Y. and J.S.-T.; Software, S.B.; Supervision, M.P.-O., E.Y. and J.S.-T.; Validation, S.B. and M.P.-O.; Writing—original draft, S.B.; Writing—review & editing, S.B., M.P.-O., E.Y. and J.S.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially conducted as part of the X5GON project funded from the EU's Horizon 2020 research programme grant No 761758. This work is also supported by the European Commission-funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437), the EPSRC Fellowship titled "Task-Based Information Retrieval" (grant EP/P024289/1), and under the AT2030 Programme. The AT2030 programme is funded by aid from the UK government and led by the Global Disability Innovation Hub.

Institutional Review Board Statement: Ethical review and approval were obtained from the University College London Research Ethics Committee, UK (Ref: 7311/001, approval date: 14 August 2019).

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [<https://github.com/sahanbull/PEEK-Dataset> and <https://github.com/sahanbull/VLE-Dataset>].

Acknowledgments: We would also like to thank the anonymous reviewers for their appreciation of the work and the thoughtful feedback that helped us significantly improve the presentation of this work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SDG	Sustainable Development Goal
AI	Artificial Intelligence
EDM	Educational Data Mining
ITS	Intelligent Tutoring Systems
EdRecSys	Educational Recommendation Systems
OER	Open Educational Resources
MOOC	Massively Open Online Courses
TF	Term Frequency
TFIDF	Term-Frequency-Inverse Document Frequency
KT	Knowledge Tracing
IRT	Item Response Theory
KC	Knowledge Components
LDA	Latent Dirichlet Allocation
INK	Interest, Novelty, Knowledge
PINK	Popularity, Interest, Novelty, Knowledge
AMD	Advanced Micro Devices
CPU	Central Processing Unit
RAM	Random Access Memory
GB	Gigabyte

References

1. Bulathwela, S.; Pérez-Ortiz, M.; Holloway, C.; Shawe-Taylor, J. Could AI Democratise Education? Socio-Technical Imaginaries of an EdTech Revolution. In Proceedings of the NeurIPS Workshop on Machine Learning for the Developing World (ML4D), Online, 14 December 2021.
2. Bulathwela, S.; Perez-Ortiz, M.; Yilmaz, E.; Shawe-Taylor, J. Towards an Integrative Educational Recommender for Lifelong Learners. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
3. Bulathwela, S.; Verma, M.; Perez-Ortiz, M.; Yilmaz, E.; Shawe-Taylor, J. Can Population-based Engagement Improve Personalisation? A Novel Dataset and Experiments. In Proceedings of the International Conference on Educational Data Mining (EDM '22), Durham, UK, 24–27 July 2022.
4. Hlosta, M.; Krauss, C.; Verbert, K.; Bonnin, G.; Millecamp, M.; Bayer, V. Workshop on Educational Recommender Systems (EdRecSys@LAK2020). 2020. Available online: <http://events.kmi.open.ac.uk/edrecsys2020/> (accessed on 3 June 2022).
5. Yudelson, M.V.; Koedinger, K.R.; Gordon, G.J. Individualized Bayesian Knowledge Tracing Models. In Proceedings of the International Conference on Artificial Intelligence in Education, Memphis, TN, USA, 9–13 July 2013; Lane, H.C., Yacef, K., Mostow, J., Pavlik, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2013.
6. Vie, J.J.; Kashima, H. Knowledge tracing machines: Factorization machines for knowledge tracing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 750–757.
7. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.J.; Sohl-Dickstein, J. Deep Knowledge Tracing. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.
8. Kim, S.; Kim, W.; Jang, Y.; Choi, S.; Jung, H.; Kim, H. Student Knowledge Prediction for Teacher-Student Interaction. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 15560–15568.

9. Pardos, Z.A.; Jiang, W. Designing for Serendipity in a University Course Recommendation System. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK '20), Frankfurt am Main, Germany, 25–27 March 2020; pp. 350–359.
10. Bulathwela, S.; Perez-Ortiz, M.; Yilmaz, E.; Shawe-Taylor, J. TrueLearn: A Family of Bayesian Algorithms to Match Lifelong Learners to Open Educational Resources. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
11. Wang, L.; Meinel, C. Mining the Students' Learning Interest in Browsing Web-Streaming Lectures. In Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining, Honolulu, HI, USA, 1–5 April 2007. [CrossRef]
12. Jiang, W.; Pardos, Z.A.; Wei, Q. Goal-based Course Recommendation. In Proceedings of the International Conference on Learning Analytics & Knowledge, Tempe, AZ, USA, 4–8 March 2019.
13. Ahmad, N.; Bull, S. Learner Trust in Learner Model Externalisations. In Proceedings of the Conference on Artificial Intelligence in Education, Brighton, UK, 6–10 July 2009.
14. Bull, S. There are open learner models about! *IEEE Trans. Learn. Technol.* **2020**, *13*, 425–448. [CrossRef]
15. Williamson, K.; Kizilcec, R.F. Effects of Algorithmic Transparency in Bayesian Knowledge Tracing on Trust and Perceived Accuracy. In Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021), Online, 29 June 29–2 July 2021.
16. UNESCO. Open Educational Resources (OER). 2019. Available online: <https://en.unesco.org/themes/building-knowledge-societies/oer> (accessed on 1 April 2019).
17. Ramesh, A.; Goldwasser, D.; Huang, B.; Daume III, H.; Getoor, L. Learning latent engagement patterns of students in online courses. In Proceedings of the AAAI Conference on Artificial Intelligence, Quebec City, QC, Canada, 27–31 July 2014.
18. Ehlers, M.; Schuwer, R.; Janssen, B. *OER in TVET: Open Educational Resources for Skills Development*; Technical Report; UNESCO-UNEVOC International Centre for Technical and Vocational Education and Training: Bonn, Germany, 2018.
19. Sunar, A.S.; Novak, E.; Mladenec, D. Users' Learning Pathways on Cross-site Open Educational Resources. In Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020), Prague, Czech Republic, 2–4 May 2020; pp. 84–95.
20. Novak, E.; Urbančič, J.; Jenko, M. Preparing Multi-Modal Data for Natural Language Processing. In Proceedings of the Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD), Ljubljana, Slovenia, 11 October 2018.
21. Pérez Ortiz, M.; Bulathwela, S.; Dormann, C.; Verma, M.; Kreitmayer, S.; Noss, R.; Shawe-Taylor, J.; Rogers, Y.; Yilmaz, E. Watch Less and Uncover More: Could Navigation Tools Help Users Search and Explore Videos? In Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, 1–5 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 90–101. [CrossRef]
22. Perez-Ortiz, M.; Novak, E.; Bulathwela, S.; Shawe-Taylor, J. An AI-Based Learning Companion Promoting Lifelong Learning Opportunities for All. 2020. Available online: https://ircai.org/wp-content/uploads/2021/01/IRCAI_REPORT_01.pdf (accessed on 3 June 2022).
23. Kang, J.; Lee, H. Modeling user interest in social media using news media and wikipedia. *Inf. Syst.* **2017**, *65*, 52–64. [CrossRef]
24. Abel, F.; Gao, Q.; Houben, G.J.; Tao, K. Twitter-Based User Modeling for News Recommendations. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13), Beijing, China, 3–9 August 2013.
25. Safari, R.M.; Rahmani, A.M.; Alizadeh, S.H. User behavior mining on social media: A systematic literature review. *Multimed. Tools Appl.* **2019**, *78*, 33747–33804. [CrossRef]
26. Piao, G.; Breslin, J.G. Inferring user interests in microblogging social networks: A survey. *User Model. -User-Adapt. Interact.* **2018**, *28*, 277–329. [CrossRef]
27. Piao, G. Recommending Knowledge Concepts on MOOC Platforms with Meta-path-based Representation Learning. In Proceedings of the International Conference on Educational Data Mining, Virtual, 29 June–2 July 2021.
28. Selent, D.; Patikorn, T.; Heffernan, N. ASSISTments Dataset from Multiple Randomized Controlled Experiments. In Proceedings of the Third (2016) ACM Conf. on Learning @ Scale (L@S '16), Scotland, UK, 25–26 April 2016; Association for Computing Machinery: New York, NY, USA, 2016. [CrossRef]
29. Bauman, K.; Tuzhilin, A. Recommending remedial learning materials to students by filling their knowledge gaps. *MIS Q.* **2018**, *42*, 313–332. [CrossRef]
30. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. -User-Adapt. Interact.* **1994**, *4*, 253–278. [CrossRef]
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. Brank, J.; Leban, G.; Grobelnik, M. Annotating Documents with Relevant Wikipedia Concepts. In Proceedings of the Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD), Ljubljana, Slovenia, 5 October 2017.
33. Piao, G.; Breslin, J.G. Analyzing Aggregated Semantics-Enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16), Halifax, NS, Canada, 13–17 July 2016.
34. Piao, G.; Breslin, J.G. Analyzing MOOC Entries of Professionals on LinkedIn for User Modeling and Personalized MOOC Recommendations. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16), Halifax, NS, Canada, 13–17 July 2016.

35. Schenkel, R.; Broschart, A.; Hwang, S.; Theobald, M.; Weikum, G. Efficient text proximity search. In Proceedings of the International Symposium on String Processing and Information Retrieval, Orlando, FL, USA, 13–15 October 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 287–299.
36. Yu, Y.; Karlgren, J.; Bonab, H.; Clifton, A.; Tanveer, M.I.; Jones, R. Spotify at the TREC 2020 Podcasts Track: Segment Retrieval. In Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020), Gaithersburg, MD, USA, 16–20 November 2020.
37. Mahapatra, D.; Mariappan, R.; Rajan, V.; Yadav, K.; Saby, A.; Roy, S. VideoKen: Automatic Video Summarization and Course Curation to Support Learning. In Proceedings of the Companion Proceedings of the The Web Conference 2018, Lyon, France, 23–27 April 2018; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2018; pp. 239–242. [[CrossRef](#)]
38. Chen, J.; Chen, X.; Ma, L.; Jie, Z.; Chua, T.S. Temporally grounding natural sentence in video. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 162–171.
39. Verma, G.; Nalamada, T.; Harpavat, K.; Goel, P.; Mishra, A.; Srinivasan, B.V. Non-Linear Consumption of Videos Using a Sequence of Personalized Multimodal Fragments. In Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21), College Station, TX, USA, 14–17 April 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 249–259. [[CrossRef](#)]
40. Bulathwela, S.; Perez-Ortiz, M.; Novak, E.; Yilmaz, E.; Shawe-Taylor, J. PEEK: A Large Dataset of Learner Engagement with Educational Videos. In Proceedings of the RecSys Workshop on Online Recommender Systems and User Modeling (ORSUM'21), Online, 2 October 2021.
41. Guo, P.J.; Kim, J.; Rubin, R. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In Proceedings of the First ACM Conference on Learning @ Scale, Atlanta, GA, USA, 4–5 March 2014.
42. Frey, M. *Netflix Recommends: Algorithms, Film Choice, and the History of Taste*; University of California Press: Berkeley, CA, USA, 2021.
43. Smith, B.; Linden, G. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Comput.* **2017**, *21*, 12–18. [[CrossRef](#)]
44. Zarrinkalam, F.; Faralli, S.; Piao, G.; Bagheri, E. Extracting, Mining and Predicting Users' Interests from Social Media. *Found. Trends® Inf. Retr.* **2020**, *14*, 445–617. [[CrossRef](#)]
45. Dinh, X.T.; Van Pham, H. A Proposal of Deep Learning Model for Classifying User Interests on Social Networks. In Proceedings of the 4th International Conference on Machine Learning and Soft Computing, Association for Computing Machinery (ICMLSC 2020), Haiphong City, Vietnam, 17–19 January 2020. [[CrossRef](#)]
46. Syed, R.; Collins-Thompson, K. Retrieval Algorithms Optimized for Human Learning. In Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR), Tokyo, Japan, 7–11 August 2017.
47. Alvarez-Melis, D.; Saveski, M. Topic Modeling in Twitter: Aggregating Tweets by Conversations. In Proceedings of the International AAAI Conference on Web and Social Media, Virtually, 7–10 June 2021.
48. Li, W.; Saigo, H.; Tong, B.; Suzuki, E. Topic modeling for sequential documents based on hybrid inter-document topic dependency. *J. Intell. Inf. Syst.* **2021**, *56*, 435–458. [[CrossRef](#)]
49. Urdaneta-Ponte, M.C.; Mendez-Zorrilla, A.; Oleagordia-Ruiz, I. Recommendation systems for education: Systematic review. *Electronics* **2021**, *10*, 1611. [[CrossRef](#)]
50. Guruge, D.; Kadel, R.; Halder, S. The State of the Art in Methodologies of Course Recommender Systems—A Review of Recent Research. *Data* **2021**, *6*, 18. [[CrossRef](#)]
51. Herbrich, R.; Minka, T.; Graepel, T. TrueSkill(TM): A Bayesian Skill Rating System. In Proceedings of the Advances in Neural Information Processing Systems 20, Vancouver, BC, Canada, 3–6 December 2007; pp. 569–576.
52. Bishop, C.; Winn, J.; Diethe, T. *Model-Based Machine Learning*; Early Access Version. 2015. Available online: <http://www.mbmlbook.com/> (accessed on 23 May 2019).
53. Schmucker, R.; Wang, J.; Hu, S.; Mitchell, T. Assessing the Performance of Online Students—New Data, New Approaches, Improved Accuracy. *J. Educ. Data Min.* **2022**, *14*, 1–45. [[CrossRef](#)]
54. Mandalapu, V.; Gong, J.; Chen, L. Do we need to go deep? knowledge tracing with big data. *arXiv* **2021**, arXiv:2101.08349.
55. Pelánek, R.; Papoušek, J.; Řihák, J.; Stanislav, V.; Nižnan, J. Elo-based learner modeling for the adaptive practice of facts. *User Model. -User-Adapt. Interact.* **2017**, *27*, 89–118. [[CrossRef](#)]
56. Bonafini, F.; Chae, C.; Park, E.; Jablow, K. How much does student engagement with videos and forums in a MOOC affect their achievement? *Online Learn. J.* **2017**, *21*. Available online: <https://www.learntechlib.org/p/183772/> (accessed on 12 January 2022). [[CrossRef](#)]
57. Lan, A.S.; Brinton, C.G.; Yang, T.Y.; Chiang, M. Behavior-Based Latent Variable Model for Learner Engagement. In Proceedings of the International Conference on Educational Data Mining, Wuhan, China, 25–28 June 2017.
58. Bulathwela, S.; Perez-Ortiz, M.; Lipani, A.; Yilmaz, E.; Shawe-Taylor, J. Predicting Engagement in Video Lectures. In Proceedings of the International Conference on Educational Data Mining (EDM '20), Virtual, 10–13 July 2020.
59. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: New York, NY, USA, 2014. [[CrossRef](#)]
60. Inigo, M.; Jameson, J.; Kozak, K.; Lanzetta, M.; Sonier, K. Combining Probabilities with “And” and “Or”. 2021. Available online: [https://math.libretexts.org/Bookshelves/Applied_Mathematics/Book%3A_College_Mathematics_for_Everyday_Life_\(Inigo_et_al\)/03%3A_Probability/3.02%3A_Combining_Probabilities_with_And_and_Or](https://math.libretexts.org/Bookshelves/Applied_Mathematics/Book%3A_College_Mathematics_for_Everyday_Life_(Inigo_et_al)/03%3A_Probability/3.02%3A_Combining_Probabilities_with_And_and_Or) (accessed on 20 May 2021).

61. Kuncheva, L.I. Ensemble Methods. In *Combining Pattern Classifiers*; John Wiley & Sons, Ltd.: New York, NY, USA, 2014; Chapter 6, pp. 186–229. [CrossRef]
62. Pardos, Z.A.; Gowda, S.M.; Baker, R.S.; Heffernan, N.T. The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *SIGKDD Explor. Newsl.* **2012**, *13*, 37–44. [CrossRef]
63. Shah, T.; Olson, L.; Sharma, A.; Patel, N. Explainable Knowledge Tracing Models for Big Data: Is Ensembling an Answer? *arXiv* **2020**, arXiv:2011.05285.
64. Burke, R. Hybrid recommender systems: Survey and experiments. *User Model. -User-Adapt. Interact.* **2002**, *12*, 331–370. [CrossRef]
65. Wu, S.; Rizoio, M.; Xie, L. Beyond Views: Measuring and Predicting Engagement in Online Videos. In Proceedings of the Twelfth International Conference on Web and Social Media, Stanford, CA, USA, 25–28 June 2018.
66. Covington, P.; Adams, J.; Sargin, E. Deep Neural Networks for YouTube Recommendations. In Proceedings of the ACM Conference on Recommender Systems, Boston, MA, USA, 15 September 2016.
67. Jannach, D.; Lerche, L.; Zanker, M. Recommending Based on Implicit Feedback. In *Social Information Access*; Springer: Berlin/Heidelberg, Germany, 2018.
68. Frey, B.J.; MacKay, D. A Revolution: Belief Propagation in Graphs with Cycles. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 30 November–5 December 1998; Jordan, M., Kearns, M., Solla, S., Eds.; MIT Press: Cambridge, MA, USA, 1998; Volume 10.
69. Minka, T. *Divergence Measures and Message Passing*; Technical Report MSR-TR-2005-173; Microsoft Research: Redmond, WA, USA, 2005.
70. Graepel, T.; Candela, J.Q.; Borchert, T.; Herbrich, R. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel, 21–24 June 2010.
71. Zaharia, M.; Chowdhury, M.; Das, T.; Dave, A.; Ma, J.; McCauly, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), San Jose, CA, USA, 25–27 April 2012; pp. 15–28.
72. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113. [CrossRef]
73. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Independent Publishers: Chicago, IL, USA, 2022.
74. Bull, S.; Kay, J. Metacognition and open learner models. In Proceedings of the 3rd Workshop on Meta-Cognition and Self-Regulated Learning in Educational Technologies, at ITS2008, 2008; pp. 7–20. Available online: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.4070&rep=rep1&type=pdf> (accessed on 12 January 2022).
75. Hooshyar, D.; Pedaste, M.; Saks, K.; Leijen, Ä.; Bardone, E.; Wang, M. Open learner models in supporting self-regulated learning in higher education: A systematic literature review. *Comput. Educ.* **2020**, *154*, 103878. [CrossRef]
76. Margetis, G.; Ntoa, S.; Antona, M.; Stephanidis, C. Human-Centered design of artificial intelligence. In *Handbook of Human Factors and Ergonomics*; John Wiley & Sons: New York, NY, USA, 2021; pp. 1085–1106.
77. Slater, S.; Baker, R.; Ocumpaugh, J.; Inventado, P.; Scupelli, P.; Heffernan, N. Semantic Features of Math Problems: Relationships to Student Learning and Engagement. In Proceedings of the International Conference on Educational Data Mining, Raleigh, NC, USA, 29 June 2016.
78. Ponzani, M.; Ferragina, P.; Chakrabarti, S. On Computing Entity Relatedness in Wikipedia, with Applications. *Knowl.-Based Syst.* **2020**, *188*, 105051. [CrossRef]
79. Liang, S. Collaborative, dynamic and diversified user profiling. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33.