Can Population-based Engagement Improve Personalisation? A Novel Dataset and Experiments

Sahan Bulathwela[1], Meghana Verma[2], María Pérez-Ortiz[1], Emine Yilmaz[1] and John Shawe-Taylor[1]
[1] Centre for Artificial Intelligence, University College London, UK
[2] Indian Institute of Technology Bombay, India
m.bulathwela@ucl.ac.uk

ABSTRACT

This work explores how population-based engagement prediction can address cold-start at scale in large learning resource collections. The paper introduces i) VLE, a novel dataset that consists of content and video based features extracted from publicly available scientific video lectures coupled with implicit and explicit signals related to learner engagement, ii) two standard tasks related to predicting and ranking context-agnostic engagement in video lectures with preliminary baselines and iii) a set of experiments that validate the usefulness of the proposed dataset. Our experimental results indicate that the newly proposed VLE dataset leads to building context-agnostic engagement prediction models that are significantly performant than ones based on previous datasets, mainly attributing to the increase of training examples. VLE dataset's suitability in building models towards Computer Science/ Artificial Intelligence education focused on e-learning/ MOOC use-cases is also evidenced. Further experiments in combining the built model with a personalising algorithm show promising improvements in addressing the cold-start problem encountered in educational recommenders. This is the largest and most diverse publicly available dataset to our knowledge that deals with learner engagement prediction tasks. The dataset, helper tools, descriptive statistics and example code snippets are available publicly.

Keywords

Population-based Engagement, Cold-start, Educational Recommender, Personalised Education, AI in Education

1. INTRODUCTION

With the growth of Open Educational Resources (OER) [38, 6, 32] and Massively Open Online Courses (MOOC) [34, 23], large educational repositories need scalable tools to understand the engagement potential of newly added materials [14]. While *contextualised engagement* can be defined as learner's engagement driven by the context at a given time in their learning path (e.g., learning needs/goals, knowledge state etc.), *context-agnostic engagement* aims to capture patterns and features associated with engagement that instead are applicable to an entire learner population rather than individual contexts of specific learners [7]. Put simply, context-agnostic engagement is concerned with the features that generally make an educational material engaging.

While many datasets capturing contextual engagement of learners exist, our contribution: **Video Lecture Engagement (VLE)**, focusing on context-agnostic engagement, is a novel dataset that presents around 12,000 peer-reviewed scientific videos constructed from a popular OER repository and contains a variety of lecture types ranging from scientific talks and expert panels to MOOC-like lectures. VLE provides textual and video-specific features extracted from the lecture transcripts, together with Wikipedia topics covered in the video (via entity linking) and user engagement labels (both explicit and implicit) for each video. In educational recommenders, VLE dataset helps solving both i) *user cold-start*, where new users join the system and we may not have enough information about their context and ii) *item cold-start*, where new educational content is released, for which we may not have user engagement data yet and thus an engagement predictive model would be necessary. While utility of context-agnostic engagement models for cold start is previously demonstrated [7], VLE is the largest dataset for building such models. This work is aimed not at replacing personalised recommendation but to complement it by addressing the cold-start problem. While VLE dataset is a major contribution of this work, several additional experimental results make up the overall contribution. These results demonstrate the usefulness of VLE dataset via answering a set of critical research questions.

2. RELATED WORK

The majority of work in Intelligent Tutoring Systems (ITS) and Educational Recommendation Systems (EduRecSys) revolve around contextual learner engagement [26, 9]. While many explore the connection between engagement and learning gains [2, 20, 28], public datasets in this realm are hard to come by. MOOC platforms such as edX [23] and Khan Academy [29] harvest valuable data created in an *in-the-wild* setting, yet this data is gated within course owners and consortia [21] (or heavily anonymised) due to its proprietary nature. However, with the boom of online education, it is greatly imperative that such datasets are democratised so that under-researched areas such as context-agnostic

S. Bulathwela, M. Verma, M. P. Ortiz, E. Yilmaz, and J. Shawe-Taylor. Can population-based engagement improve personalisation? A novel dataset and experiments. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 414–421, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853185

(population-based) engagement can be widely understood to push the frontiers of AI-supported online education. Study of context-agnostic engagement of video lectures so far has been mainly qualitative, deriving guidelines such as keeping videos short and in parts [23, 3]. While these findings are useful in content creation stage, they have little use in moderating the mammoth of materials already circulating in the Internet. Our work proposes building predictive models that can be utilised post-creation for scalable quality assurance and content recommendation.

2.1 Related Datasets

Many video engagement prediction works revolve around YouTube[16, 24] and uses platform specific features (e.g. channel reputation, category etc.) exclusively. While large-scale datasets for this task is published, these datasets include general-purpose videos (largely entertainment related) rather than educational videos [40]. Some features proposed in these works share similarity to our proposal (such as duration, language and topic features). However, no content-based features (based on transcript) relating to understandability and presentation are used, making the methods hard to generalise outside of YouTube. Educational Information Retrieval [36, 15] and Wikipedia page quality prediction [19, 39] has been attempted using features such as text style, readability, structure, network, recency and review data. Publicly available Wikipedia article quality dataset [19] with human annotated (explicit) labels is used to tackle the latter task although implicit labels are not included in this dataset. Similar datasets are available for automated essay scoring [37]. But, none of these datasets fill the lack of datasets for predicting engagement of educational videos.

In the educational context, measuring learner engagement using multi-modal data such as brain waves [41], facial expressions [25] etc. are conducted in controlled environments and lacks "in-the-wild" signals [22]. Large public datasets and competitions also relate to students answering questions in-the-wild (e.g. ASSISTments [30] or multiple choice questions [13]), contrary to the proposed VLE dataset, lacks implicit feedback relating to learners acquiring skills/knowledge. More relevant datasets studying population-based engagement revolves around MOOCs. Studying approximately 800 videos from edX platform, Guo et al. [23] manually processed and provided a qualitative analysis of engagement, with a few features being relatively subjective and difficult to automate. A similar work [35] takes 22 edX videos, extracts cross-modal features and manually annotates their quality with no focus on learner engagement. Neither dataset is publicly available. MOOCCube is a recently released dataset that contains a spectrum of details relating to MOOC interactions [42]. Although large, the video watch logs in MOOCCube come from 190,000 users contrary to over 1.1 Million users of VLE. The dataset is also not tested for engagement prediction by its publishers raising uncertainty in its promise.

Our prior work [7] explored the possibility of building context-agnostic engagement models using implicit watch time-based labels and showed that these models can address the cold-start problem. We identify this contribution most relevant to the proposed dataset. Our prior work publishes a dataset of 4,000 lectures with labels coming from 150,000 learners. VLE, expands this dataset with 3 times as many videos with engagement signals generated by 7 times as many learners. The new dataset also restricts itself to English lectures to refine relevance of the proposed features. Additionally, VLE introduces a new set of Wikipedia-based topical features.

3. VLE DATASET

VLE dataset is created using the aggregated video lectures consumption data coming from VideoLectures.Net (VLN). These videos are recorded when researchers are presenting their work at peer-reviewed conferences. Lectures are thus reviewed and material is controlled for correctness of knowledge. The collection consists of scientific talks and tutorials that are mainly geared towards university level learners. In that aspect, many videos in the dataset are stylistically similar to conventional MOOC lectures. VLE dataset provides a set of features together with labels based on subjective assessment metrics such as star ratings and view count. We believe that this dataset enables understanding the connection between content features and the collective engageability of learners with an educational video.

3.1 Feature Extraction

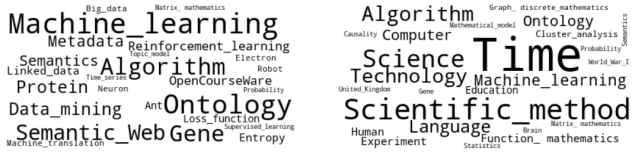
The video metadata and transcriptions are transformed into i) content-based textual features, ii) Wikipedia topic-based features and iii) video-specific features. Majority of the extracted features are cross modal (e.g. books, websites and audios) and are easily automatable.

Content-based Features. Based on prior proposals [7], we extract *Word Count* [39], *Title Word Count* and *Document Entropy* [1], language style features [18], *Preposition Rate*, *Auxiliary Rate*, *To Be Rate*, *Conjunction Rate*, *Normalisation Rate*, *Pronoun Rate*, readability related *Easiness (FK Easiness)*[18] and vocabulary related *Stop-word Presence Rate*, *Stop-word Coverage Rate* [1, 31]. To represent *Freshness* of lectures (recency), we calculate the number of days between January 01, 1960 and the video published date to use it as a proxy for recency of the lecture [7].

Wikipedia-based Features. We use Wikifier [4] to extract topical features capturing topic authority and coverage.

The *top-5 authoritative topic URLs* and *top-5 PageRank scores* features represent the Topic Authority feature vertical. Wikifier [4] produces a PageRank score [5] that indicates the marginal authoritativeness of a Wikipedia concept among all Wikipedia concepts associated with a lecture. It is noteworthy that *authority* of a learning resource entails author, organisation and content authority [11]. The proposed features represent content authority.

The *top-5 covered topic URLs* and *top-5 cosine similarity scores* features represent *Topic Coverage* feature vertical. The cosine similarity score $cos(s_{tr}, c)$ between the *Term Frequency-Inverse Document Frequency (TF-IDF)* representations of the lecture transcript s_{tr} and the Wikipedia page of concept c is also an output from the Wikifier. These features are used as a proxy for topic coverage.



(i) Most Authoratative (*PageRank based*) (ii) Most Covered (*Cosine based*)

Figure 1: WordClouds summarising the distribution of top 25 (i) most authoritative and (ii) most covered Wikipedia topics in the dataset. Note that Data Science and Computer Science related topics are most dominant topics.

These four feature sets create 20 distinct feature columns. Figure 1 presents two word clouds that show the 25 most authoritative and covered topics in the VLE dataset and show that there are distinct differences between topic presence in the two feature sets. The authoritative topics are topics that are highly connected to other topics in the lecture whereas covered topics have high textual overlap between lecture transcript and Wikipedia concept page.

Video-specific Features. We identify a set of easily automatable, prior proposed [7] video specific features. *Lecture Duration, Is Chunked, Lecture Type*[23], *Silence Period Rate (SPR)* and *Speaker Speed* [7] are calculated based on prior work . *Lecture Duration* is reported in seconds. *Is Chunked* is a binary feature which indicates if a lecture has multiple parts. *Lecture type* value is derived from the metadata. This diverse dataset contains many different types of videos ranging from presentations, panels to tutorials.

3.2 Labels

Three main types of quantification of engagement labels are presented in the dataset.

Explicit Ratings and Popularity. *Mean Star Rating* based on a scale of 1-5 (5 being best) for each video is provided. Ratings are paired with the number of ratings used to compute the mean. The proposed dataset has 2,127 ratings (almost 2x than [7]). Missing ratings are labelled with `-1`. Capturing popularity, the total number of views, named *View Count*, for each video as of February 1, 2021 is provided.

Watch Time/Engagement. Many labels in this dataset are based on *watch time* [16]. Normalised Engagement Time (NET) is computed for each video as it has been proposed as the gold standard for learner engagement [23]. The `Median of NET (MNET)` and `Average of NET (ANET)` is calculated from NET. To have the MNET and ANET labels in the range $[0, 1]$, we set the upper bound to 1, deriving Saturated MNET (*SMNET*) and Saturated ANET (*SANET*) that are included in the dataset. The standard deviation of `NET` (*Std of Engagement*) is also reported, together with the *Number of User Sessions* used for calculating `MNET` and `ANET`. These measures allow understanding stability of the centres published. The set of individual NET values are also published to allow future researchers to exploit the true distribution of values.

3.3 Preserving Anonymity and Ethics

Users of VLN repository formally agree that all user generated content is available for research. We further anonymise and aggregate interactions to ensure privacy. Aggregated results are only published in lectures with atleast 5 user views to preserve k-anonymity [17]. The presenters and authors of videos in VLN also formally agree to the video, presentation slides and supplementary material to be published under an open license and can be used for educational and research purposes. Still, a regime of techniques outlined below are used to preserve presenter anonymity in order to avoid having unanticipated effects on lecturer's reputation by associating implicit learner engagement values to their content. Rarely occurring *Lecture Type* values are grouped together to create the `other` category. Life Sciences, Physics, Technology, Mathematics, Computer Science, Data Science and Computers categories are grouped as `stem` and all other categories as `misc` category. Rounding is used with *Freshness* and *Lecture Duration* to the nearest 10 days and 10 seconds respectively. Gaussian white noise (10%) is added to *Title Word Count* feature and rounded to the nearest integer.

VLN repository is concentrated with videos about Computer Science (see Figure 1), a subject area with gender imbalance in both audience and presenters. We avoided using feature classes that could potentially reflect gender characteristics to improve neutrality of the dataset (and models). Visual features (facial features, emotions...) and audio features (pitch, tone...) that may actively/passively embed gender is avoided. We focused primarily on features that reflect informational content. Where video specific features are used, generic features such as "speaker speed" that are unlikely to be correlated to gender or age are used.

3.4 Final Dataset

The final dataset contains 11,548 lectures across 21 subjects (eg. Computer Science, Philosophy, etc. with a majority from AI and Computer Science) that are published between September 1, 1999 and December 31, 2020. The engagement labels are created from events of over 1.1 Million users logged between December 01, 2016 and February 01, 2020. The collection of videos span various video lengths with the duration distribution having two modes at approx. 2000s (33 mins) and 4000s (1hr) time points which align with typical lengths of research talks and presentations. The mean word count of the videos is 5347.9. The video lecture collection uses on average 93.9 learners per video when calculating engagement centres. The dataset, helper tools and example code snippets are available publicly[1].

3.5 Supported Tasks

Scalable Quality Assurance and Educational Recommendation are two key downstream applications of context-agnostic engagement prediction. We establish two main tasks, which we mainly focus on in this paper, that can be objectively addressed using the VLE dataset. These are:

- **Task 1:** Predicting context-agnostic (population-based) engagement of video lectures
- **Task 2:** Ranking of video lectures based on engagement

[1] `https://github.com/sahanbull/VLE-Dataset`

Other Tasks. Beyond the proposed tasks, this dataset is suitable for, not limiting to, several tasks such as i) understanding influential features for engagement prediction and ii) understanding the strengths and weaknesses of different implicit/explicit labels, that have been investigated in our prior work with similar datasets [7, 33]. The textual and topical representations, with the use of unsupervised approaches can be used in a range of tasks from understanding meaningful hidden patterns within clusters of videos (e.g. talks vs. lectures vs. tutorials) to deducing the structure of knowledge based on how topics co-occur within videos.

3.6 Evaluating Performance

We identify *Root Mean Squared Error (RMSE)* as a suitable metric for evaluating Task 1. Measuring RMSE against the original labels published with the datasets will allow different works to be compared fairly. With reference to Task 2, we identify *Spearman's Rank Order Correlation Coefficient (SROCC)* as a suitable metric. SROCC is suitable for comparing between ranking models that create global rankings (e.g. point-wise rankers).

We use 5-fold cross validation to evaluate model performance with tasks 1 and 2. The folds are released together with the dataset, to allow to facilitate fair comparison and reproducability. 5-fold cross validation allows reporting the *confidence intervals (1.96 × Standard Error)* of the performance estimate, which we include in Table 1.

4. BASELINES AND EXPERIMENTS

Through our experiments, we aim to answer research questions relevant to the supported tasks (in section 3.5) and other facets that further demonstrate the utility of this dataset. The main research questions of interest are:

- **RQ1:** Does the newly constructed VLE dataset lead to training more performant prediction models?
- **RQ2:** How does the larger quantity of training data affect predictive performance?
- **RQ3:** Is the model useful for modelling engagement with Computer Science materials?
- **RQ4:** Is this dataset useful for modelling engagement in E-Learning lectures and MOOC videos?
- **RQ5:** Does context-agnostic engagement prediction help in the cold-start scenario?

Our previous work [7] demonstrated *Random Forest (RF)* model obtains best performance among linear and non-linear models in similar datasets. Therefore, we use the RF model to benchmark the new VLE dataset for Tasks 1 and 2 described earlier.

4.1 Labels and Features for Baseline Models

SMNET label is used as the target variable for both tasks. Preliminary investigations indicated that SMNET label follows a Log-Normal distribution, motivating us to use a log transformation on the SMNET values before training the models. Empirical results further confirmed that this step improves the final performance of the models. We undo this transformation for computing $RMSE$ while this transformation doesn't affect $SROCC$.

All the features outlined as the content-based and video-based sections in section 3.1 are included in the baseline models. The models are trained with three different feature sets in an incremental fashion:

1. *Content-based*: Features extracted from lecture metadata and the transcript-based textual features.
2. *+ Wiki-based*: Content-based + 2 Wikipedia-based features (Top 1 Most Authoritative Topic URL and Most Covered Topic URL).
3. *+ Video-based*: Content-based + Wiki-based + Video-specific features.

However, due to the large amount of topics in the Wikipedia-based feature groups, we restrict to the top 1 authoritative and covered topic features where they are encoded as binary categorical variables. Practitioners are encouraged to try further encoding of the topic variables, as it will likely have a positive impact on the performance.

4.2 Experiments

Addressing RQ1, the RF models are trained with the three proposed feature sets using 5-fold cross validation with the prior proposed, smaller `4k` dataset [7] and the newly proposed VLE dataset (`12k`). This setup allows identifying how performance gains are achieved through i) adding each new group of features and ii) adding new observations. Follow on experiments addressing RQ2 and RQ3 are run using hold-out validation technique where fold 5 is held out. We experiment by using varying proportions of training data in RQ2 to train the model. When selecting training data, random sampling is used All the trained models in RQ2 are evaluated using the same held out test set.

To validate RQ4, we partition the entire dataset into i) tutorial videos (`vtt` lecture type) and ii) all other videos, as test and train data respectively. However, tutorials conducted in conferences significantly vary from e-learning videos geared for MOOCs. To address this mismatch, we further identified 1,035 videos (among the tutorials) that exclusively belong to the Open Course Ware Consortium (OCWC)[2]. OCWC exclusively contains university lectures intended for teaching and devises different MOOC production techniques such as classroom, talking head and power point methods [23]. In RQ5, we train the model using all non-tutorial lectures and test the engagement prediction/ranking performance on i) OCWC videos (`ocw`), ii) all tutorials but OCWC (`!ocw`) and iii) all tutorials `vtt`, (entire test set). Experiments of (RQ2-4) are only done with the best performing model from Table 1 (RF model with *Content + Wiki + Video* feature group) to reduce computational cost.

To tackle RQ5, we utilise TrueLearn Novel [9] (hereby referred to as *TrueLearn*), a personalisation model that predicts learner engagement with video lectures. A key limitation of many such models is lack of information in the

[2] `http://videolectures.net/ocwc`

Table 1: RMSE and SROCC with confidence intervals for the engagement prediction (Task 1) and lecture ranking (Task 2) using the Random Forests model with both 4k [7] and 12k (Our VLE) datasets. Better performance highlighted in bold.

	RMSE with Task 1		*SROCC* with Task 2	
Feature set	4k	12k *(Ours)*	4k	12k *(Ours)*
Content-based	.1801±.006	**.1170±.006**	.6190±.011	**.7504±.013**
+ Wiki-based	.1798±.007	**.1178±.006**	.6251±.014	**.7505±.013**
+ Video-specific	.1728±.007	**.1098±.007**	.6758±.020	**.7832±.009**

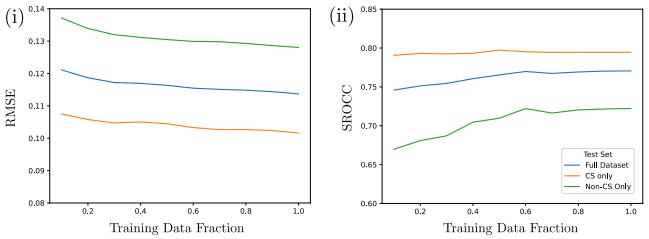


Figure 2: Predictive performance for (i) engagement prediction and (ii) lecture ranking tasks with varying proportions of randomly sampled training data. The test set performance for full test dataset (Blue) and subsets of test dataset that consists of CS lectures only (Orange) and Non-CS lectures only (Green) are also reported

Table 2: Performance for OpenCourseWare (ocw), Non-OpenCourseWare (!ocw) tutorial and All tutorial (vtt) videos for engagement prediction and lecture ranking tasks. Better performance per task is highlighted in bold.

	ocw	!ocw	vtt	From Table 1
RMSE with Task 1	.0539	**.0404**	.0406	.1098
SROCC with Task 2	**.9485**	.9209	.9223	.7832

early stages of the user session (*cold-start* problem) to make accurate predictions. In this experiment, we created a hybrid recommender that combines the TrueLearn model with the proposed context-agnostic engagement prediction model (hereby referred to as *TrueLearn*$_{++}$). For simplicity, we use "switching" [12] in TrueLearn$_{++}$, where the pre-trained context-agnostic model makes the prediction on the *first* event of each user (where the personalistion model has no information) and switches to TrueLearn that can exploit user history. PEEK dataset [8], with more than 20,000 learner sessions, is used for the experiment where the context agnostic model is trained using lectures not present in the PEEK test data. Then the predictive performance on the PEEK test data using TrueLearn (The baseline) and TrueLearn$_{++}$ (the hybrid model) is measured using Accuracy, Precision, Recall and F1-Score. A learner-wise, one-tailed paired t-test is used for statistical significance testing.

5. RESULTS AND DISCUSSION

The performance metrics observed with the RF model on Task 1 and 2 (RQ1) are outlined in Table 1. Figure 2 illustrates how the training data size impacts the i) RMSE and ii) SROCC (RQ2) as well as showing the engagement prediction performance on Computer Science (CS) videos vs. Non-CS videos (RQ3). Table 2 presents predictive performance of the model on e-learning type lectures and tutorials (RQ4). Finally, the overall performance comparison relating to the effect of *combining* the context agnostic model with personalisation models to battle cold-start problem (RQ5) is reported in Table 3.

5.1 Performance Gains and Causes (RQ1-2)

Table 1 shows that the larger VLE leads to significant performance gains in both engagement prediction and video ranking tasks over the 4k dataset [7]. When using all feature sets with the RF RMSE on Task 1 drops by 41% while SROCC on Task 2 jumps by 15%. The labels in VLE dataset have more statistical stability as the centres are computed using more data points collected over a wider time window. Using more feature groups also leads to improved performance. Results for VLE dataset in Table 1 shows this trend where best performance is evidenced with all the features (*+ Video-specific* group) used. However, using cross-modal content-based features (*Content-based + Wiki-based*) alone leads to substantial performance. This result indicates that engagement prediction is still feasible only depending on easy to automate, cross modal features. Wikification, used in generating Wiki-based features, also operates in web-scale[3]. While Table 1 results don't show a significant bump when Wiki-based features are added, we believe that this is due to the simplicity of features used with much room for more sophisticated features (e.g. semantic relatedness between Wikipedia topics [10]) that will lead to performance gains.

Figure 2 confirms that the increase of training data improves performance in both tasks. RMSE continues to shrink in Figure 2(i) while SROCC in Figure 2(ii) tells a different story where improvements saturate at 60%. This suggests that improving ranks gets significantly harder around 5,500 training examples (≈ 60% of training set).

5.2 Relevance to AI/CS Education (RQ3) and E-learning Scenarios (RQ4)

Figure 2 shows that VLE dataset is suitable for training models for CS-only lectures leading to test set RMSE of ≈ .1 and SROCC of ≈ .8. This may be due to i) the higher diversity (significant differences between subjects) of lectures within the non-CS group and ii) the majority of CS/AI (e.g. Machine Learning, Ontology, Semantic Web etc.) related videos being present within the VLE dataset (see Figure 1). Table 2 shows strong evidence that the models trained with VLE dataset generalise really well for engagement modelling in e-learning type videos created for course teaching amid the dataset containing many different video types. The models trained are much better at engagement prediction and ranking of tutorial-like videos than general scientific talks. Having tested with lectures that have been recorded using different MOOC video production techniques, the high performance obtained on ocw lectures confirms that VLE dataset can be highly effective in building context-agnostic engagement models for e-learning and MOOC systems.

5.3 Addressing the Cold-Start Problem (RQ5)

Table 3 shows that simply combining the context-agnostic engagement prediction with TrueLearn Novel algorithm (together becoming TrueLearn$_{++}$) can lead to significant improvements in accuracy and precision. The same table also

[3] http://wikifier.org

Table 3: Average test set performance for Accuracy (Acc.), Precision (Prec.), Recall (Rec.) and F1-Score (F1) in predicting *first* event and *all* events. The more performant value is highlighted in bold. The metrics where the proposed model that outperform the baseline counterpart in the PEEK dataset ($p < 0.01$ in a one-tailed paired t-test) are marked with $\cdot^{(*)}$.

	Predicting *first* event				Predicting *all* events			
Model	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Truelearn	44.21	44.21	**100.00**	**61.32**	62.69	57.54	**81.88**	**64.98**
Truelearn$_{++}$	**56.09**$^{(*)}$	**50.32**$^{(*)}$	53.58	51.90	**63.51**$^{(*)}$	**57.91**$^{(*)}$	79.13	64.39

shows that the drop of overall F1 score can be attributed to the comparatively steeper drop of Recall Score. Inspecting left part of Table 3 sheds more light into where this steep drop of recall occurs. This is, the TrueLearn model always predicts positive engagement for the first event of the user. At the *first* event, the recall of TrueLearn being 1.0 while the accuracy and precision being the same depicts this fact. TrueLearn predicts positive in event 1 of each user because the model has no information to base the prediction on [9]. However, the scenario is different in TrueLearn$_{++}$ in the first event as the model has additional information. Both accuracy and precision of predictions in first event of the learner population significantly improves. 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 additional information is 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 first event of learners based on lecture content features enable significantly improving prediction accuracy and precision of TrueLearn$_{++}$. It is also noteworthy that our experiment, for the sake of simplicity, uses a rule that could be significantly improved further, e.g. using weights of the probabilities of both population-based and personalised models at the beginning of a user session (also known as stacking [12]), where the weight of population-based engagement decreases as we gather more information about the user.

5.4 Opportunities and Limitations

The VLE dataset is one of the biggest datasets for modelling context-agnostic engagement of educational videos (15× than [23] and 3× than [7]). This unlocks potential to apply complex model families (e.g. deep learning) with the potential to periodically expand the dataset to further push the frountiers of research. The Wiki-based features open up limitless possibilities as many sophisticated feature sets can be built and experimented. Due to the connectivity to Wikipedia, both its content and link structures can be exploited to invent meaningful, yet interpretable features. A further step can enable other data structures such as knowledge bases (e.g Wikidata) and category trees. to be used for feature creation. As the VLE dataset captures how content features relate to engagement, this dataset can be used to solidify our understanding on how to create engaging learning materials [27, 23, 3, 7].

There also exists limitations. VLE dataset is largely comprised of Computer Science and Data Science materials (Figure 1) that are delivered all in English. While this is an opportunity for AI and Computer Science education, results in Figure 2 also shows that this fact leads to comparatively *less fruitful* non-CS results. The dataset and its features are also not suitable for non-English video collections. Amid its size, the dataset still lacks variety of materials in topical and lingual sense. As pointed out in section 3.3, we have taken some measures to restrict the feature set to what we believe to be more neutral features that do not discriminate gender or age. However, since we do not have access to gender information in the data collected, it is impossible to test and guarantee that VLE dataset doesn't carry negative gender,age biases. Care should be taken when enhancing these features and there is room to do more rigorous tests to understand if any gender biases are present within the dataset. *Learner Engagement* is a loaded concept with many facets. In relation to consuming videos, many behavioural actions such as pausing, rewinding and skipping can contribute to latent engagement with a video lecture [28]. Due to the technical limitations of the platform and privacy concerns, only watch time, view and mean ratings are included in this dataset. Although watch time has been used as a representative proxy for learner engagement with videos [23, 40, 16], we acknowledge that more informative measures may lead to more complete engagement signals.

6. CONCLUSION

In order to push the frontiers of context-agnostic engagement prediction, we construct and release the VLE dataset consisting of i) content-based, ii) Wiki-based and iii) video-specific features with multiple explicit and implicit labels. Two formal tasks are established to predict engagement and rank videos with baseline models that significantly outperform predecessors. Empirically, i) improvement of performance with training data size, ii) suitability of VLE dataset for CS/AI education, iii) relevance to MOOCs and e-learning and iv) the feasibility of using context-agnostic engagement prediction to address cold-start problem in personalised (contextual) educational recommedation is demonstrated.

In retrospect of section 5.4, expanding the dataset vertically (with diverse observations) and horizontally (with novel features) is our highest priority going forward. As a future direction, much richer learner engagement signals (e.g. pauses, replays, skips etc.) can be incorporated to the dataset without compromising user privacy. When better understanding of learner engagement is gained, training examples coming from other modalities (e.g. PDF and E-books) can be added to the dataset to further widen the scope of the dataset, enabling understanding of learner engagement across different types of learning resources. The attempt to improve personalisation using population-based models in our work is barely scratching the surface. There is a lot of potential to extensively propose sophisticated and rigorously tested approaches to exploit this idea, which we will also explore in future work.

7. ACKNOWLEDGMENTS

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) and the EPSRC Fellowship titled "Task Based Information Retrieval" (grant EP/P024289/1).

8. REFERENCES

[1] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proc. of ACM Int. Conf. on Web Search and Data Mining*, 2011.

[2] F. Bonafini, C. Chae, E. Park, and K. Jablokow. How much does student engagement with videos and forums in a mooc affect their achievement? *Online Learning Journal*, 21(4), 2017.

[3] C. J. Brame. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education*, 15(4), 2016.

[4] J. Brank, G. Leban, and M. Grobelnik. Annotating documents with relevant wikipedia concepts. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)*, 2017.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of Int. Conf. on World Wide Web*, 1998.

[6] S. Bulathwela, S. Kreitmayer, and M. Pérez-Ortiz. What's in it for me? augmenting recommended learning resources with navigable annotations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, IUI '20, 2020.

[7] S. Bulathwela, M. Perez-Ortiz, A. Lipani, E. Yilmaz, and J. Shawe-Taylor. Predicting engagement in video lectures. In *Proc. of Int. Conf. on Educational Data Mining*, EDM '20, 2020.

[8] S. Bulathwela, M. Perez-Ortiz, E. Novak, E. Yilmaz, and J. Shawe-Taylor. Peek: A large dataset of learner engagement with educational videos, 2021.

[9] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources. In *AAAI Conference on Artificial Intelligence*, AAAI '20, 2020.

[10] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. Semantic truelearn: Using semantic knowledge graphs in recommendation systems. In *Proc. of First Int. Workshop on Joint Use of Probabilistic Graphical Models and Ontology*, PGMOnto '21, 2021.

[11] S. Bulathwela, E. Yilmaz, and J. Shawe-Taylor. Towards Automatic, Scalable Quality Assurance in Open Education. In *Workshop on AI and the United Nations SDGs at Int. Joint Conf. on Artificial Intelligence*, 2019.

[12] R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[13] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, C. Bae, B. Kim, and J. Heo. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*, pages 69–73. Springer, 2020.

[14] K. Clements and J. Pawlowski. User-oriented quality for oer: understanding teachers' views on re-use, quality, and trust. *Journal of Computer Assisted Learning*, 28(1), 2012.

[15] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proc. of Conf. on Information and Knowledge Management*, 2011.

[16] P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *Proc. of ACM Conf. on Recommender Systems*, 2016.

[17] N. Craswell, D. Campos, B. Mitra, E. Yilmaz, and B. Billerbeck. Orcas: 20 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 2983–2989, New York, NY, USA, 2020. Association for Computing Machinery.

[18] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. A general multiview framework for assessing the quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and Technology*, 2017.

[19] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality*, 2(3), Dec. 2011.

[20] D. Davis, G. Chen, C. Hauff, and G. Houben. Activating learning at scale: A review of innovations in online learning strategies. *Comput. Educ.*, 125:327–344, 2018.

[21] D. Davis, D. Seaton, C. Hauff, and G. Houben. Toward large-scale learning design: categorizing course designs in service of supporting learning outcomes. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale, London, UK, June 26-28, 2018*, pages 4:1–4:10, 2018.

[22] M. A. Dewan, M. Murshed, and F. Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1, 2019.

[23] P. J. Guo, J. Kim, and R. Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proc. of the First ACM Conf. on Learning @ Scale*, 2014.

[24] M. Horta Ribeiro and R. West. Youniverse: Large-scale channel and video metadata from english-speaking youtube. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1), 2021.

[25] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018.

[26] S. Kim, W. Kim, Y. Jang, S. Choi, H. Jung, and H. Kim. Student knowledge prediction for teacher-student interaction. *Proceedings of the AAAI Conference on Artificial Intelligence*,

35(17):15560–15568, 2021.

[27] M. Kurdi, N. Albadi, and S. Mishra. "think before you upload": an in-depth analysis of unavailable videos on youtube. *Social Network Analysis and Mining*, 11(1):1–21, 2021.

[28] A. S. Lan, C. G. Brinton, T.-Y. Yang, and M. Chiang. Behavior-based latent variable model for learner engagement. In *Proc. of Int. Conf. on Educational Data Mining*, 2017.

[29] Z. MacHardy and Z. A. Pardos. Evaluating the relevance of educational videos using bkt and big data. In *Proc. of Int. Conf. on Educational Data Mining*, 2015.

[30] M. Mendicino, L. Razzaq, and N. T. Heffernan. A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education*, 41(3):331–359, 2009.

[31] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. of Int. Conf. on World Wide Web*, 2006.

[32] M. Perez-Ortiz, C. Dormann, Y. Rogers, S. Bulathwela, S. Kreitmayer, E. Yilmaz, R. Noss, and J. Shawe-Taylor. X5learn: A personalised learning companion at the intersection of ai and hci. In *26th International Conference on Intelligent User Interfaces - Companion*, IUI '21 Companion. Association for Computing Machinery, 2021.

[33] M. Perez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019.

[34] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Proc. of AAAI Conference on Artificial Intelligence*, 2014.

[35] J. Shi, C. Otto, A. Hoppe, P. Holtz, and R. Ewerth. Investigating correlations of automatically extracted multimodal features and lecture video quality. In *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information*, SALMM '19, page 11–19, New York, NY, USA, 2019. Association for Computing Machinery.

[36] R. Syed and K. Collins-Thompson. Optimizing search results for human learning goals. *Inf. Retr. J.*, 20(5):506–523, 2017.

[37] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*, 2016.

[38] UNESCO. Open educational resources (oer). `https://en.unesco.org/themes/building-knowledge-societies/oer`, 2021. Accessed: 2021-04-01.

[39] M. Warncke-Wang, D. Cosley, and J. Riedl. Tell me more: An actionable quality model for wikipedia. In *Proc. of Int. Symposium on Open Collaboration*, 2013.

[40] S. Wu, M. Rizoiu, and L. Xie. Beyond views: Measuring and predicting engagement in online videos. In *Proc. of the Twelfth Int. Conf. on Web and Social Media*, 2018.

[41] F. Xu, L. Wu, K. P. Thai, C. Hsu, W. Wang, and R. Tong. MUTLA: A large-scale dataset for multimodal teaching and learning analytics. *CoRR*, abs/1910.06078, 2019.

[42] J. Yu, G. Luo, T. Xiao, Q. Zhong, Y. Wang, W. Feng, J. Luo, C. Wang, L. Hou, J. Li, et al. Mooccube: a large-scale data repository for nlp applications in moocs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3135–3142, 2020.