X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number:	761758
Project Acronym:	X5GON
Project title:	Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date:	2017-09-01 to 2020-08-31
Project Duration:	36 months
Deliverable Title:	D1.6 – Report on Selected Models and Content Representations
Lead beneficiary:	UCL
Type:	Report
Dissemination level:	Public
Due Date (in months):	36 (August 2020)
Date:	31-December-2020
Status (Draft/Final):	Final
Authors:	Sahan Bulathwela, Maria Perez-Ortiz, E. S. V. Ranawaka, R. I. P. B. B. Siriwardana, G. A. K. Y. Ganepola, Emine Yilmaz and John Shawe-Taylor
Contact persons:	Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor

Revision

Date	Lead author(s)	Commments
01/12/2020	Sahan Bulathwela Maria Perez-Ortiz, Emine Yilmaz and John Shawe-Taylor	Initial Draft
08/12/2020	E. S. V. Ranawaka R. I. P. B. B. Siriwardana and G. A. K. Y. Ganepola	Added Chapters on Language Detection and Duplicate Detection
14/12/2020	Alfons Juan	Internal Review
28/12/2020	Sahan Bulathwela	Final Version

Contents

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

List of Figures

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

List of Tables

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

Abstract

Multiple content representation models have been proposed in deliverable D1.4 – Advanced Content Representations [1] towards improving automatic language detection, duplicate detection and improved personalisation of educational materials, all tasks which are aimed towards improving the quality of the X5GON Open Educational Resource (OER) database and its services. The evaluation methodologies utilised for model selection are published in deliverable D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs [2] with the novel datasets that were constructed to support evaluation. This report presents the results obtained through evaluation. An ensemble of `fasttext` and `cld2` was selected as the final model to classify mono/multi-lingual documents with languages due to their reported accuracy of 97.5% and 95.53% on the language dataset. The selected duplicate detection model (that uses both TF-based and Wikifier-based content representation) obtained 99% precision. Gradient boosting machine model was chosen for context-agnostic engagement prediction based on its superior pairwise ranking accuracy score.

In the context of advancing the personlisation model, Semantic TrueLearn algorithm which uses `W2V` semantic relatedness metric is selected due to its superior F1-score 83.7%. This model also obtains superior accuracy and precision scores.

[1] https://www.x5gon.org/science/deliverables/

[2] https://www.x5gon.org/science/deliverables/

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

1 Introduction

This report outlines the detailed results obtained from the evaluation of content representation and quality models that were developed in the X5GON project. Content representation models that related to quality assurance of educational materials [1], language detection, duplicate detection and personalisation of learning materials [2] have been advanced in the latter part of the project. The advanced content representation models that have been proposed are outlined extensively in deliverable *D1.4 – Advanced Content Representations*. Deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs* describes the appropriate evaluation methodologies utilised the novel datasets that have been made available to the public in the process of creating datasets to evaluate the models.

This report presents in detail, the results obtained during the evaluation and the conclusions derived from the observed results. Through the evaluation, the most suitable models are selected and discussed.

1.1 Chapter Overview

The main contents in this report are broken into two chapters. Chapter 2 presents the results, discussions, conclusions and potential future directions relating to 3 quality related content representations, namely, (i) language detection model, (ii) duplicate detection model and (iii) context-agnostic engagement prediction model. Chapter 4 presents the results and discussion relating to *Semantic TrueLearn Novel* model is presented. Finally, the conclusions are derived from the results and future directions are proposed.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

2 Selected Quality Models

In this chapter, we focus on evaluating and selecting 3 content representation models that are related to improving the overall quality of X5GON database.

1. Language Detection Model
2. Duplication Detection Model
3. Context-Agnostic Engagement Prediction Model

2.1 Language Detection Model

The objective of building a language detection model is to use existing language detection model to build a reliable, fast language detection system that works with multiple languages. As per the deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs*, we follow a two-step evaluation process where the models are tested with a dataset filled with monolingual documents. Then, we expand our analysis to a dataset with bilingual documents that are synthetically generated. Classification accuracy and computational time is used to measure performance.

2.1.1 Data and Models

Based on the languages of documents available in the X5GON database [3], 8 languages were chosen for the study. Namely, they are, German(`de`), Dutch(`nl`), English(`en`), Slovene(`sl`), Slovak(`sk`), French(`fr`), Italian(`it`) and Spanish(`es`). All selected languages contain 1000 data points each.

Different popular language detection libraries were identified and bench marked. Libraries such as `spacy`, `langdetect`, `cld2` etc. were evaluated on both mono- and bilingual settings. The full list of libararies can be found in deliverable *D1.4 – Advanced Content Representations*.

2.1.2 Results and Discussion

Monoligual Document Dataset Initially, the models were tested with the monolingual dataset. The results for the prediction accuracy and computational time are presented in figures 1 and 2.

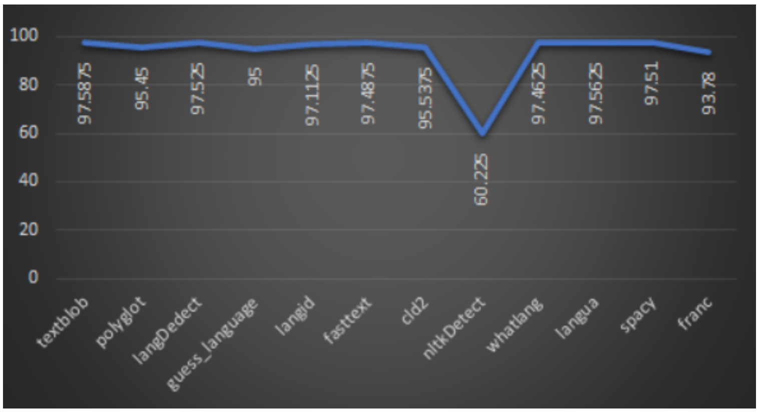


Figure 1: Predictive accuracy of different language detection models with the monolingual document dataset. Higher values closer to 100 are better.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

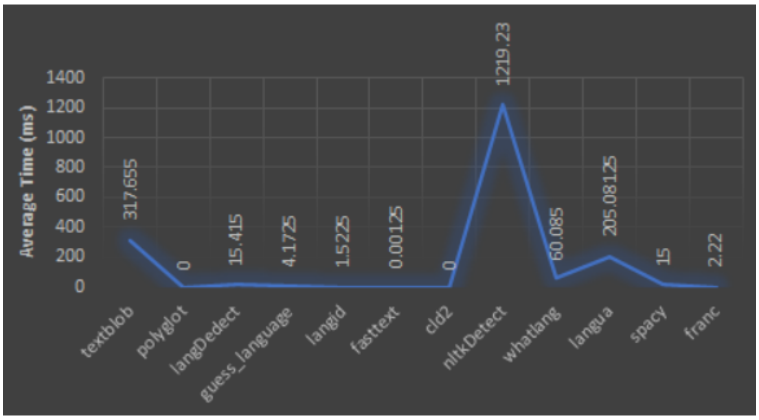


Figure 2: Time incurred in milliseconds (ms) of different language detection models with the monolingual document dataset. Lower values closer to 0 are better.

Bilingual Document Dataset In the bilingual case, there are many different ways accuracy can be interpreted for predictive performance. Three of these metrics are as follows:

1. *Pair-wise Accuracy*: Observation considered to be predicted accurately if,
   (a) Both languages (full pair) are predicted correctly
   (b) The order of dominance is predicted correctly
2. *Dominant Language*: Observation considered to be predicted accurately if the dominant language in the observation is predicted correctly. The second dominant language may be mis-classified or not predicted at all.
3. *Both Languages*: Observations considered to be predicted accurately if the two languages are predicted correctly. The order may be mis-classified.

As per the deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs*, predictive accuracy is tested for varying proportions of most dominant and secondary languages. The accuracy of the model for 20:80 proportion is presented in figure 3. The accuracy of the model for 30:70 proportion is presented in figure 4. The accuracy of the models for 40:60 proportion and 50:50 proportion are presented in figures 5 and 6 respectively. The computational time costs for different language detection libraries on the bilingual datasets is presented in figure 7.

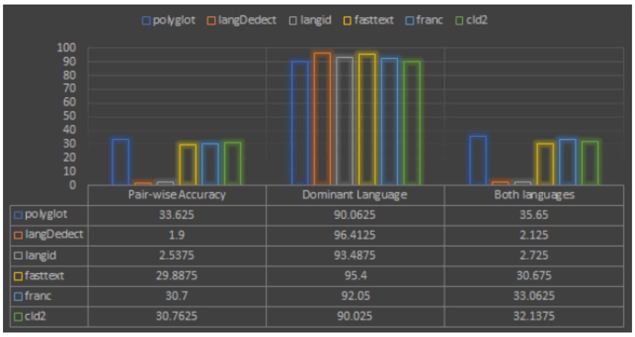
Discussion Figures 1 and 2 demonstrates `nltkdetect` performs the worst in both accuracy and time. In terms of accuracy `textblob`, `langua`,`langdetect`, `spacy` and `fasttext` demonstrate similar performance while `polyglot`, `cld2`, `fasttext`, `langid` and `franc` indicate to be computationally efficient in detecting the dominant language in a document.

For the bilingual document dataset, the following libraries show promise based on the performance and ability to detect multiple languages efficiently: `polyglot`, `langdetect`, `langid`, `fasttext` and `franc`. Figures 3 ...6 show that the ability to detect the dominant language in each library seems to deteriorate as the split proportion between the two languages reaches 50-50 while the pairwise accuracy and the accuracy in detecting both languages improve.

`langdetect` performs better than `fasttext` in detecting the dominant language in terms of accuracy yet in terms of time `fasttext` performs significantly better. `polyglot` and `cld2` libraries have the

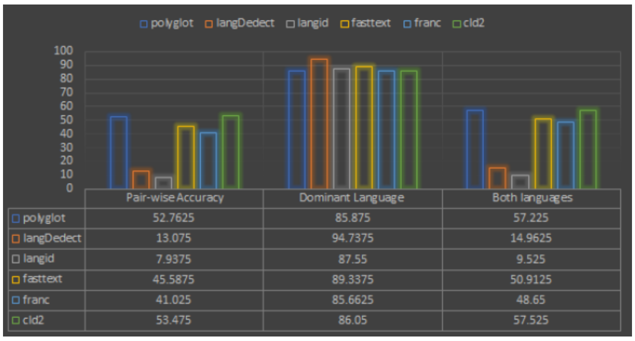
Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

highest pairwise accuracy performance yet they don't perform as well as `langdetect` and `fasttext` in detecting the dominant language.



	Pair-wise Accuracy	Dominant Language	Both languages
□ polyglot	33.625	90.0625	35.65
□ langDedect	1.9	96.4125	2.125
□ langid	2.5375	93.4875	2.725
□ fasttext	29.8875	95.4	30.675
□ franc	30.7	92.05	33.0625
□ cld2	30.7625	90.025	32.1375

Figure 3: Predictive accuracy on the bilingual dataset where the languages are mixed 20:80 in favour of the most dominant language. Higher values closer to 100 are better.



	Pair-wise Accuracy	Dominant Language	Both languages
□ polyglot	52.7625	85.875	57.225
□ langDedect	13.075	94.7375	14.9625
□ langid	7.9375	87.55	9.525
□ fasttext	45.5875	89.3375	50.9125
□ franc	41.025	85.6625	48.65
□ cld2	53.475	86.05	57.525

Figure 4: Predictive accuracy on the bilingual dataset where the languages are mixed 30:70 in favour of the most dominant language. Higher values closer to 100 are better.

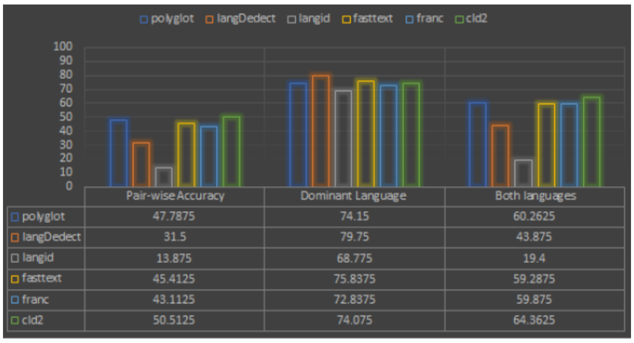
2.1.3 Selected Language Detection Model

As the original X5GON database may include documents written both a single language and multiple languages, using a single library for language detection may result in loss of accuracy. Thus a strategy of using an ensemble classifier is proposed. Within the ensemble, the issue of detecting and ranking of multiple languages of a document is proposed to be solved in two steps.

1. Detecting the dominant language
2. Detecting multiple languages (if document consist of multiple languages)

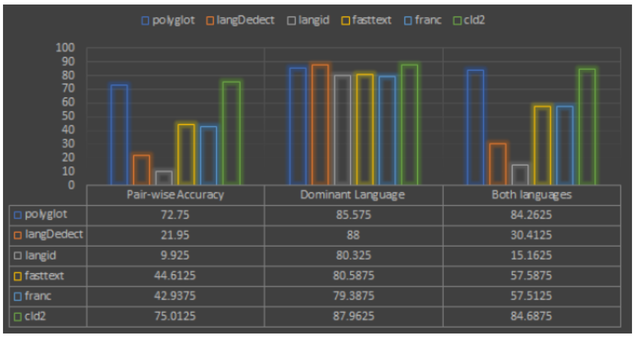
This is to be achieved through a pipeline which uses `fasttext` and `cld2` on the documents in the mentioned order. Selecting `cld2` over `polyglot` is motivated by the less restrictive re-usability capabilities of the Apache license [4, 5] and better accuracy and computation time.

Performance of Ensemble Model Once we have identified the ensemble model, we test the predictive performance of this model using the dataset we have created. We use the monolingual dataset here as the majority of the items in X5GON database are monolingual [3]. Table 1 shows the accuracy of predicting the chosen 8 languages in the dataset. We also construct the confusion matrix



	Pair-wise Accuracy	Dominant Language	Both languages
polyglot	47.7875	74.15	60.2625
langDedect	31.5	79.75	43.875
langid	13.875	68.775	19.4
fasttext	45.4125	75.8375	59.2875
franc	43.1125	72.8375	59.875
cld2	50.5125	74.075	64.3625

Figure 5: Predictive accuracy on the bilingual dataset where the languages are mixed 40:60 in favour of the most dominant language. Higher values closer to 100 are better.



	Pair-wise Accuracy	Dominant Language	Both languages
polyglot	72.75	85.575	84.2625
langDedect	21.95	88	30.4125
langid	9.925	80.325	15.1625
fasttext	44.6125	80.5875	57.5875
franc	42.9375	79.3875	57.5125
cld2	75.0125	87.9625	84.6875

Figure 6: Predictive accuracy on the bilingual dataset where the languages are mixed 50:50 in favour of the most dominant language. Higher values closer to 100 are better.

(presented in figure 8) for above stated dataset to test out the performance of the implemented language detection tool proposed above. The confusion matrix points out that there is high tendency for all the considered major European languages to have a tendency to be classified as *English*. Although the mis-classification rate is not significantly high, this is a phenomenon that should be addressed in future versions of the language detector.

2.2 Duplication Detection Model

Multiple approaches are useful in detecting duplicate OERs from X5GON database. The most obvious strategies involves using deterministic rules that can be used to identify duplicate materials. We identify multiple different approaches are feasible which are both deterministic and non-deterministic.

2.2.1 Deterministic Duplicate Detection Rules

We hypothesise that there are multiple rules that can deterministically identify duplicate materials. Some of these rules are based on:

1. *Content Similarity and Filenames:* if the content in two documents is identical, they are duplicates. Also, identical files tend to be named identically.

2. *URL redirects:* if the OER URL redirects to another URL in the Database, they are duplicates.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

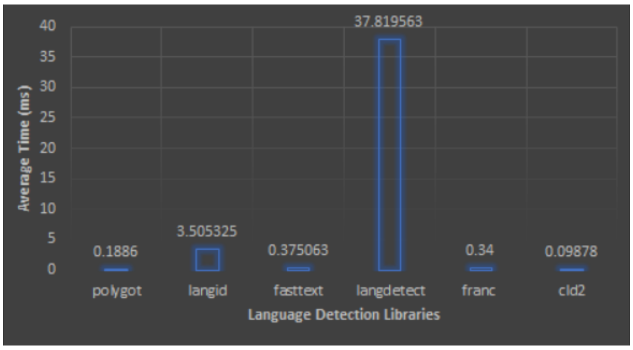


Figure 7: Time incurred in milliseconds (ms) of different language detection models with the bilingual document dataset. Lower values closer to 0 are better.

Table 1: Classification accuracy of the eight major European Languages.

Language Code	Language	Accuracy
en	English	0.999
nl	Dutch	0.946
sk	Slovak	0.906
es	Spanish	0.948
sl	Slovene	0.936
it	Italian	0.938
de	German	0.975
fr	French	0.972

Content Similarity and Filenames As the first step for duplication detection we checked for exact duplicates in the current database using the value to value comparison of documents.

Then we looked at how the file hash and the filename correlates with duplication. Figure 9 shows how the number of files compare with number of unique file hashes and filenames.

URL Redirects Altogether, it could be observed that 711 out of the duplicates are redirects that account to 327 unique values. These URLs are from two domains. These URL redirects come from two OER repositories as pointed in table 2.

2.2.2 Content Representation based Similarity Detection

For two documents to be duplicates, having the same set of words is not sufficient. In other words, if two bags of words were created for two documents, just those two bags being similar is not sufficient to determine whether the documents were similar.We devise Term Frequency based and Wikification based content representations to capture token level and topic level signal from documents. For more information about the content representations, we direct the reader to deliverable *D1.5 – Evaluation*

Table 2: URL redirects detected from OER repositories.

OER Repository	Total Duplicated URLs	Distinct Resources
ocw.mit.edu	34	1
cnx.org	677	326

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

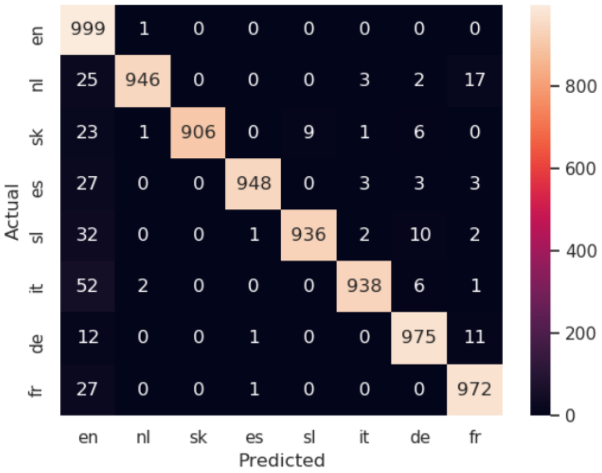


Figure 8: Confusion matrix for classifying major European languages using the proposed language detection system. The lighter cells indicate high overlap vs. darker cells indicate low overlap.

Table 3: Predictive Performance of duplicate prediction models based on TF and Wikifier Content Representations.

TF Threshold	Wikifier Threshold	Selected	Duplicates	Not Duplicates	Precision
High	High	100	99	1	**0.990**
High	Low	50	0	50	0.000
Low	High	35	9	26	0.257

Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs.

TF Similarity vs. Word Count It is intuitive that two candidate documents for duplication should be of similar document length. We plot the word count difference between the document pairs and their TF cosine similarity to investigate this relationship. Figure 10 shows the relationship between TF similarity and word count difference.

TF Similarity and Wikifier Similarity As the two content representation strategies (TF and Wikifier) has unique strengths in representing token level and topic level features respectively, a combination of the featuresets is experimented with. The results from this experiment are outlined in table 3.

2.2.3 Results and Discussion

Among 8,542 exact duplicate materials that were detected, 3,230 distinct values (textual content) exist; which implies that 5,312 documents can be labelled as duplicates. Out of the total of 3230 documents 46.32% of them have the same hashes in the OER materials that are categorised into them. 52.93% have the same filenames in all OER materials categorised into them. 82 categories have duplicate OER materials across multiple domains.

Figure 9 shows that the *same hash* is a good metric for duplication. This analysis also determines whether the same filename could be used as an alternate metric for duplication. However, if the file hash is the same, the contents of the file will also be identical. Therefore, considering either the file

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

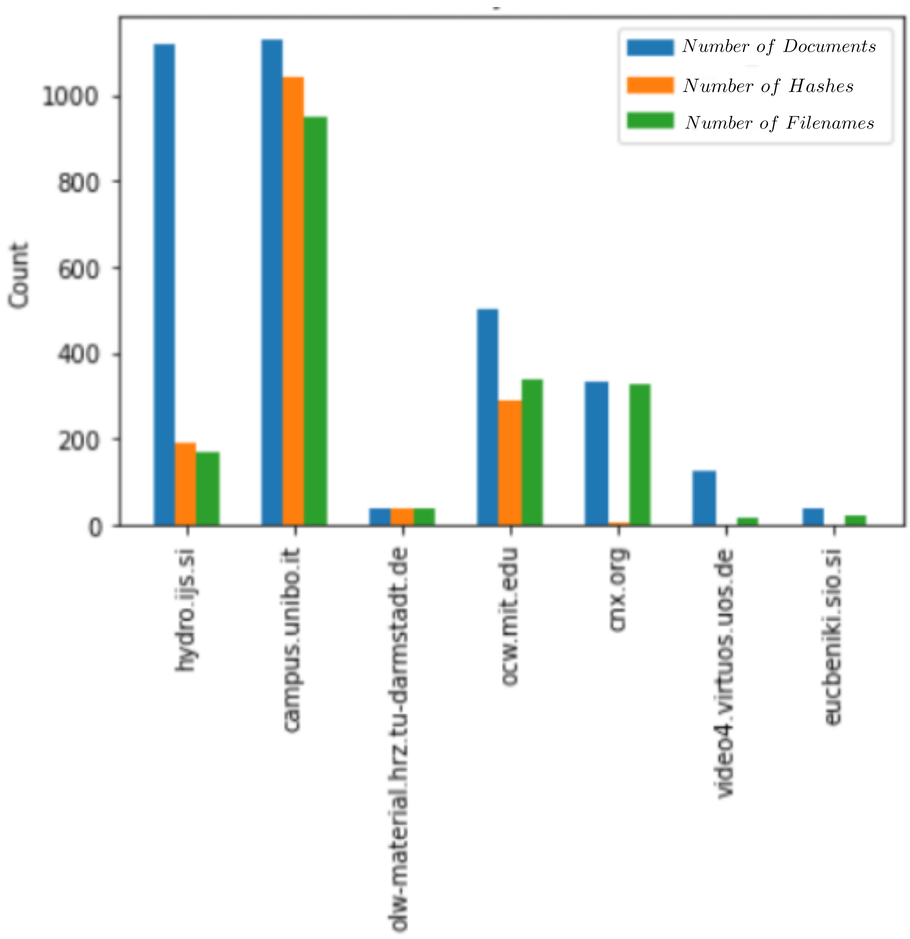


Figure 9: How the number of unique file hashes, filenames among the documents change based on the OER repository.

contents or the hashes alone is sufficient for the comparison. However, it is hard to consider the *same filename* as a good metric as different files can be named identically.

Out of the two domains in table 2, `cnx.org` reports:

"...All community-created content within CNX will remain accessible in a read-only state for two years. We want to make sure that any current users have time to arrange alternate plans for their content. After this transition period, all of CNX's community-created content will be housed at Internet Archive (www.archive.org), where it will be freely viewable and downloadable. We chose Internet Archive, not only for its reliability but also because of their mission, "to provide universal access to all knowledge," aligns with ours!

This change doesn't affect the OpenStax-published textbooks – our library of books will continue to be available on OpenStax.org. In fact, we're devoting more resources to developing and enhancing our online reading experience and we're planning to publish more titles in the future. We're excited about the impact these changes are already having on students, like our recently released highlighting and note-taking capability, available for most of our titles! Try it out here. ..."

Thus, all `cnx.org` URLs redirect to `openstax.org` domain. Although these `cnx.org` redirects to distinct URLs the values are duplicates. Thus for `cnx.org` and other domains, duplicates can be also determined by evaluating only the content similarity.

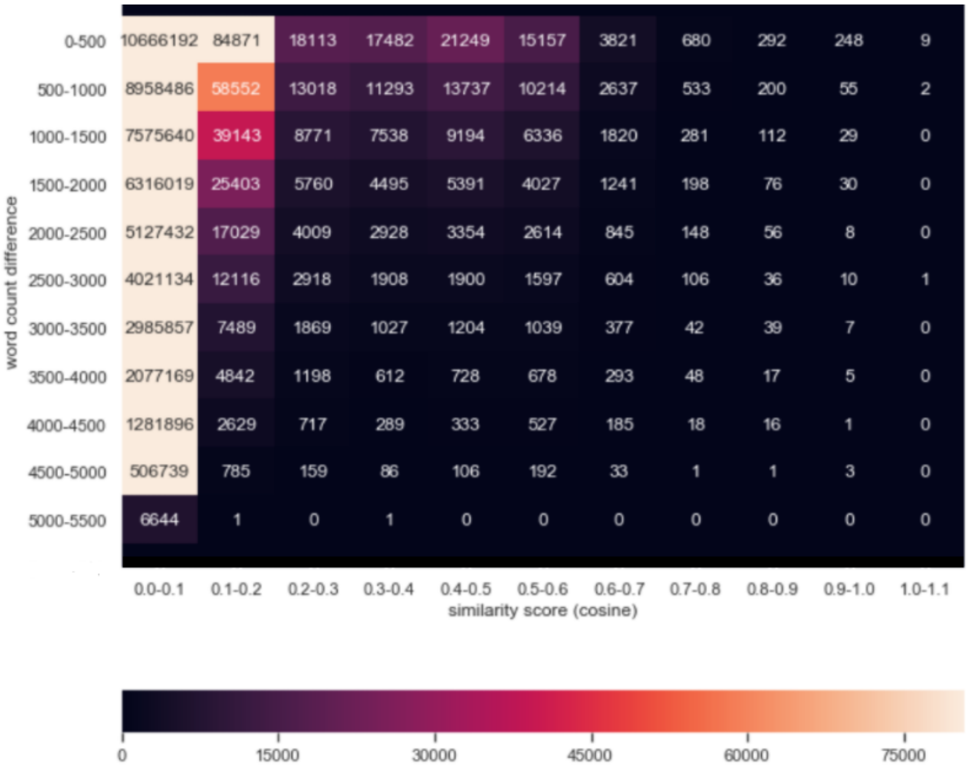


Figure 10: How the content similarity between document pairs changes when the word count difference is changing.

2.3 Context-Agnostic Engagement Prediction Model

As discussed in deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs*, we identify 2 out of the 7 tasks identified in relation to context-agnostic engagement prediction to be evaluated. The predictive performance of the models are evaluated using Gradient Boosting Machines (GBM) [6] and Random Forests (RF) model [1] due to their superior performance in similar tasks.

2.3.1 Feature Sets

The models are trained with three different feature sets in an incremental fashion:

1. *Content-based*: Features extracted from lecture metadata and the textual features extracted from the lecture transcript.

2. *+ Wiki-based*: In addition to the content-based features, two Wikipedia based features (most authoritative topic URL and most covered topic URL) are added to the feature set.

3. *+ Video-based*: In addition to both content-based and Wikipedia-based features, video specific features are added.

2.3.2 Results and Discussion

The results for engagement prediction task (Task 1) are reported in Table 4. Table 5 reports the performance in ranking lectures based on engagement (Task 2). It is evident that addition of Wikipedia-

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

Table 4: Test RMSE for the engagement prediction models (task 1) with standard error (lower values are better).

	RMSE	
Feature Set	GBM	RF
Content-based	.1802±.0160	**.1801±.0137**
+ Wiki-based	.1814±.0160	**.1798±.0148**
+ Video-specific	.1737±.0172	**.1728±.0160**

Table 5: Test SROCC and Pairwise Ranking Accuracy (Pairwise) for lecture ranking models (task 2) with standard error (higher values are better).

Model	GBM		RF	
Feature Set	SROCC	Pairwise	SROCC	Pairwise
Content-based	**.6241±.0291**	**.7221±.0102**	.6190±.0237	.7202±.0086
+ Wiki-based	.6245±.0339	.7224±.0115	**.6251±.0322**	**.7225±.0123**
+ Video-specific	**.6761±.0434**	**.7446±.0183**	.6758±.0458	.7446±.0197

based features and video-specific features contribute towards improving model performance across both tasks with video-specific features leading to significant gains. The results show that the RF model is consistently better at predicting lecture engagement (Table 4) whereas the GBM model dominates the performance in lecture ranking (Table 5) although these two models belong to the ensemble learning family.

This dataset provides us with the opportunity to understand context-agnostic engagement with a unique type of video lectures, specifically, scientific videos. Although the results in tables 4 and 5 show that adding Video-specific features leads to consistent improvements of predictive performance, it is evident that the cross-modal content-based features alone lead to substantial amount of predictive performance in comparison to the gains by adding modality-specific features. This is a good indication that easy-to-compute, cross-modal features alone are sufficient to build a system that can predict context-agnostic engagement of video lectures to a satisfactory degree.

The results also indicate that there is no significant gain in performance by adding the Wikipedia features. However, we believe that this is due to the simplicity of the Wiki features used in constructing the baselines. A portfolio of more informative features could be built using the Wikipedia information provided with the dataset.

2.3.3 Limitations and Opportunities

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 [7]. Analysing facial expressions and affective states is another alternative approach to representing engagement [8]. However, due to the technical limitations of the platform and privacy concerns, only watch time, number of views and mean ratings are included in this dataset. Although watch time has been used as a representative proxy for learner engagement with videos [9, 10], we acknowledge that more informative measures may lead to more complete and reliable engagement signals.

Although this is the case, there are numerous opportunities that are presented by this dataset. It provides the opportunity to understand engagement with scientific videos and to what extent the engagement dynamics align/differ with other types of educational videos. In addition to the summarised engagement signals, the individual user engagement signals are provided with the dataset.

This data will allow researchers to better understand the engagement distribution and apply more creative techniques to flesh out the engagement signals.

2.4 Conclusions and Future Directions

To address the need of context-agnostic engagement prediction that can improve scalable quality assurance and recommendations systems in education, we have constructed and published a novel dataset with a wide range of features for over 4000 scientific video lectures. The dataset consists of a diverse set of lectures belonging to multiple languages, knowledge areas and lecture types with features that are content-based, Wikipedia-based and video specific. In the spirit of improving engagement prediction in video lectures, we establish two main tasks, (i) predicting context-agnostic engagement of video lectures and (ii) ranking video lectures based on engagement, together with 7 auxiliary tasks that can be addressed with this dataset. Ensemble learning methods tend to perform well in this task, leading to introducing two baseline models for the two main tasks. The promising performance of the models with the dataset demonstrates the possibility of building machine learning models to predict engagement in video lectures.

We plan several lines of future work relating to improving the limitations of the current version of the dataset (and therefore the potential tasks it can be used for). This entails both horizontal and vertical expansion of the dataset. Horizontal expansions relates to introducing new features. More content-based features can be computed by exploiting the semantic graph constructed with the Wikipedia topics [11]. A wider range of features that capture textual, audio-visual and presentation slides related patterns will be constructed [12]. Computer vision based features for videos and processing visual information in educational material (slides in videos) can be provided to improve modality-specific feature sets. Vertical expansions of the dataset relate to adding new observations. Adding more video lectures coming from multiple sources such as YouTube would widen the diversity of data. Following the reflections from section 2.3.3, the possibility of including more learner engagement related signals (e.g.: pauses, replays, skips, etc.) will be explored in the subsequent version of the dataset, without compromising learner privacy. As more understanding of engagement with other modalities (such as PDFs and e-Books) is gained, it is possible to add more observations from diverse modalities to widen the horizons of the dataset and improve understanding of engagement with different modalities of educational material. Additional features with more diverse observations and representations may unlock the possibility of experimenting with more sophisticated deep learning and multi-task learning models. We will also connect the dataset to learners' personalised data through our future work in order to support building personalised tasks and making the connection to population-based engagement, which has been suggested in previous work as an important step towards building integrative educational recommender systems [13].

3 Selected Content Representation Model

In this section, we report how the most suitable content representation is chosen. As pointed out in deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs*, we follow a two step process where we (i) select the most suitable weight combination for the KC ranking in the former step and (ii) identify the most suitable learner model in the latter step.

3.1 Evaluation Criteria

A sequential experimental design is employed, where engagement of fragment t is predicted using fragments 1 to $t-1$. We also use a one hold-out validation approach for hyper-parameter tuning where hyper-parameters are learned on 70% of the learners and the model is evaluated on the remaining 30% with the best hyper-parameter combination. Since engagement is binary, predictions for each fragment can be assembled into a confusion matrix, from which we compute well-known binary classification metrics such as accuracy, precision, recall and F1-measure. We average these metrics per learner and weight each learner according to their amount of activity in the system. We use F1-measure to rank the models as we are interested in improving both precision and recall. The results obtained on the test dataset are reported in table 7. The evaluation criteria and the offline dataset used for the evaluation is reported extensively in deliverable *D1.5 – Evaluation Methodologies for Content Representation Models and Release of Datasets for Measuring Quality of OERs* where more details can be found.

3.2 Knowledge Tracing Vs. TrueLearn Novel

In the context building personalised learning systems while preserving interpretability, Knowledge Tracing (KT) model [14] is the most sought after model. Educational data mining researchers also tend to desire transparent models as student models [15] which has led to continuous evolution of KT model through individualisation [16], incorporating additional components such as interventions [17] and many other improvements. However, it is demonstrated that the proposed TrueLearn model outperforms KT model in terms of predictive performance while preserving interpretability [2, 18]. The dataset we have consists of a majority of users who have very short sessions. Based on this information, we hypothesise that TrueLearn algorithm might be more performant on shorter sessions in comparison to the KT model. We construct a plot to validate this hypothesis. The plot is presented in figure 11.

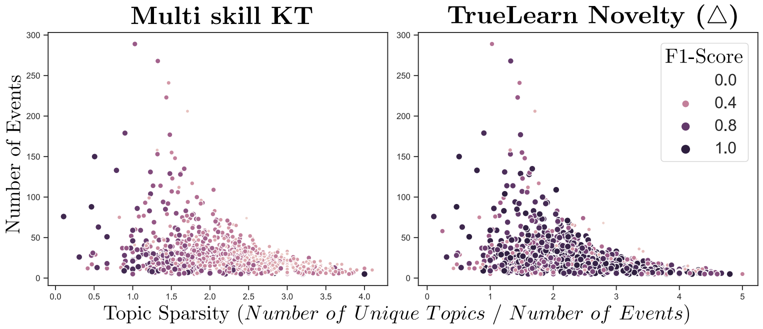


Figure 11: F1 score of each learner with their associated topic sparsity (x-axis) and number of events (y-axis). Each data point represents a learner. Colours represent F1-Score.

3.3 Weighting for KC Ranking

In the first step, we identify the best weight combination for KC ranking that is compatible with TrueLearn Novel algorithm [2]. To gain a detailed understanding of the TrueLearn Novel algorithm, please refer to the deliverable *D1.3 – Initial Content Representations* [18]. We follow a grid search to identify the most suitable combination of weights. The results from this experiment are reported in table 6.

Table 6: Predictive performance of TrueLearn Novel algorithm with different weight combinations for $W_{PageRank}$ and W_{Cos}. The different configurations of weight combinations are evaluated using Accuracy, Precision , Recall and F1 Score. The most performant value and the next best value are highlighted in **bold** and *italic* faces respectively.

$W_{PageRank}$	W_{Cos}	Accuracy	Precision	Recall	F1-Score
0.0	1.0	0.7844	**0.7852**	0.8619	0.8177
0.1	0.9	0.7955	0.7826	0.9000	0.8318
0.2	0.8	0.7921	0.7766	**0.9322**	0.8377
0.3	0.7	0.7965	0.7785	0.9154	0.8344
0.4	0.6	0.7962	0.7782	0.9170	0.8348
0.5	0.5	0.8006	0.7812	0.9276	0.8403
0.6	0.4	0.8009	0.7822	0.9292	0.8415
0.7	0.3	0.7999	0.7814	0.9294	*0.8410*
0.8	0.2	*0.8016*	0.7825	*0.9314*	**0.8424**
0.9	0.1	**0.8023**	*0.7833*	0.9277	0.8417
1.0	0.0	0.8003	0.7830	0.9191	0.8385

3.4 Adding Semantic Relatedness Information

As the second step, we attempt to investigate if leveraging semantic relatedness information from Wikipedia will improve TrueLearn algorithm. Many competing definitions of semantic relatedness (SR) exist [11] and we aim to identity the most suitable one for Semantic TrueLearn by evaluating different SR metrics. The results are presented in table 7.

3.5 Discussion

Figure 11 compares how F1 score changes for individual learners with respect to number of events and topic sparsity. It is evident that KT model struggles with learners that have high topic sparsity. The definition of *topic sparsity* here is *number of unique events per interaction event in the session.* This means that the learners that encounter a diverse set of topics in a few number of events will have a high topic sparsity value. We can observe from Figure11 that many users end up with a F1-score of 0 when topic sparsity is greater than 4 where as there are many users who obtain a high F1-score with topic sparsity between 4 and 5.

Table 6 shows that the classification *accuracy* tend to increase with increasing $W_{PageRank}$. Recall and precision doesn't show a dominant trend. In the context of finding the suitable weight combination for KC ranking, $W_{PageRank}$ of 0.8 and W_{cos} of 0.2 seems to be the ideal combination of weights. F1-Score is the harmonic mean of precision and recall which represents the best of both metrics. A high

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

Table 7: The different configurations of Semantic TrueLearn Novel algorithm are evaluated using Accuracy, Precision , Recall and F1 Score. The most performant value and the next best value are highlighted in **bold** and *italic* faces respectively. The Semantic TrueLearn algorithms that outperform baseline model in terms of F1 score are underlined.

Algorithm	SR Metric	Accuracy	Precision	Recall	F1-Score
TrueLearn Novel	-	0.7807	0.7667	0.9476	0.8348
Semantic TrueLearn Novel	`M&W`	0.7830	0.7701	0.9469	*0.8364*
	`W2V`	**0.7837**	**0.7714**	0.9467	**0.8370**
	`PMI`	0.7813	0.7682	*0.9480*	0.8355
	`LM`	0.7763	0.7605	**0.9507**	0.8322
	`Jaccard`	0.7763	0.7605	**0.9507**	0.8322
	`CP`	0.7773	0.7621	**0.9507**	0.8330
	`BA`	*0.7827*	*0.7704*	0.9469	*0.8364*

$W_{PageRank}$ here means that the authority of topics is much more important in KC ranking which will lead to a better result with TrueLearn algorithm.

The performance gain obtained by using semantic relatedness is presented in table 7. The overall results show that predictive performance gains can be achieved by incorporating semantic relatedness information between the Wikipedia topics which leads the *Semantic TrueLearn Novel* algorithm, the evolution from *TrueLearn Novel* algorithm that is proposed in deliverable *D1.3 – Initial Content Representations* [18]. It is evident from table 7 that incorporating semantic relatedness leads to improvements in overall F1 score in majority of the SR metrics beating the baseline TrueLearn algorithm. Four Semantic TrueLearn models (ones that use `M&W`, `W2V`, `PMI` and `BA` SR metrics) tend to outperform the baseline TrueLearn Novel model in terms of accuracy, precision and F1. The remainder of Models (`LM`, `Jaccard` and `CP` SR metric based models) tend to lead in recall. Entity embedding-based SR metric (`W2V`) leads to the best performing model. This is expected as neural-based semantic relatedness measures often outperform their graph-based counterparts [11].

3.6 Conclusion

Given the results, we choose $W_{PageRank}$ of 0.8 and W_{cos} of 0.2 as the most suitable combination of weights for KC ranking. We choose `W2V` embedding-based SR Metric with the Semantic TrueLearn model to improve the personalising model.

Leveraging semantic relatedness between Wikipedia topics seems a promising approach to improve predictive performance of algorithms such as TrueLearn that are built on Wikipedia ontology. The results obtained in the above experiments show that incorporating semantic relatedness information about knowledge components lead to better performance, exploiting the additional information available to it. Semantic relatedness can be truly valuable in early stages of the user session when the interaction data about the user is limited. We identify that most KCs encountered by the model in a session are highly correlated. This leads to overlapping information being propagated repeatedly which may lead to overestimation of knowledge of unseen KCs in the latter part of learner session as many correlated topics are used for the estimation. Further details with examples about this phenomenon is presented in appendix A.1.

3.6.1 Future Work

there is promise in using methods such as PageRank [19] to derive skill parameters that are uncorrelated, which is an avenue yet to be explored. Alternatively, building a hierarchical representation of knowledge [20] consisting of mutually exclusive (uncorrelated) Wikipedia concepts can be done in the future. Moreover, semantic relatedness measures are not usually built and validated with educational datasets or topics, which is a limitation. In the future, we will validate these measures with an educational dataset.

In terms of building a more holistic and integrative personalisation system, learner specific aspects that go beyond knowledge and novelty, such as content quality, learner interests have to be incorporated to the learner model [13]. The future work should focus on unifying these individual pieces together.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

A Appendix

A.1 Semantic Relatedness Between Wikipedia-based Knowledge Components

Entity Linking is devised to automatically annotate educational resources with Wikipedia-based Knowledge Components (KCs). While this approach addresses the cost intensive expert labelling bottleneck, one negative outcome of this approach is that the KCs that are assigned are likely to be highly correlated to each other. Figure 1 demonstrates two such scenarios that are found in the dataset used. What is evident here is that overlapping information between the two observed topics are being propagated to inferring the unseen topic which may lead to an overestimation of the value.

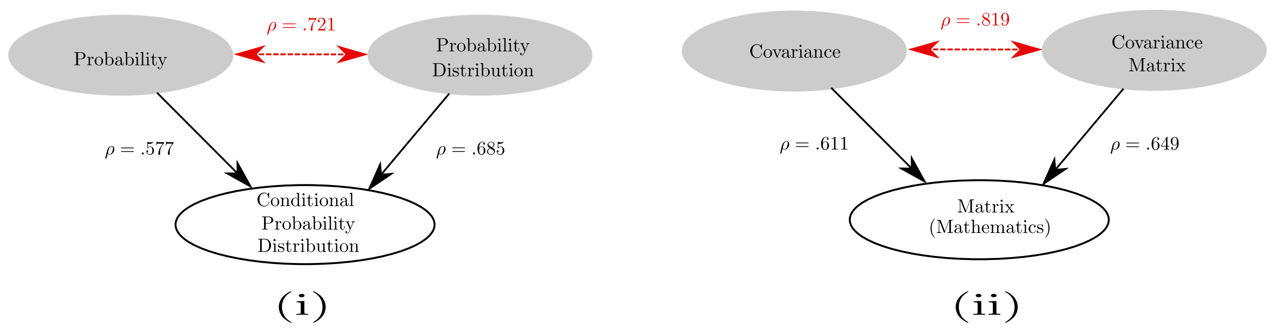


Figure 12: Two random instances from the dataset where unseen KCs (transparent nodes) are estimated with observed (shaded nodes) using Semantic Relatedness ρ. Semantic relatedness between the observed topics is shown in red dashed arrows. In (i), parameter for *Conditional Probability Distribution* is estimated using parameters of variables *Probability* and *Probability Distribution* and in (ii), *Matrix (Mathematics)* being estimated using parameters *Covariance* and *Covariance Matrix*.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

References

[1] Sahan Bulathwela, Maria Perez-Ortiz, Aldo Lipani, Emine Yilmaz, and John Shawe-Taylor. Predicting engagement in video lectures. In *Proc. of Int. Conf. on Educational Data Mining*, EDM '20, 2020.

[2] 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*, 2020.

[3] Erik Novak, Jasna Urbančič, and Miha Jenko. Preparing multi-modal data for natural language processing. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)*, 2018.

[4] Apache Softwaare Foundation. Community-led development "the apache way". `https://www.apache.org/licenses/GPL-compatibility.html`. Accessed: 2020-12-05.

[5] Michel Bauwens. Why apache defeated the gpl license: developer freedom vs. user freedom. `https://blog.p2pfoundation.net/why-apache-defeated-the-gpl-license-developer-freedom-vs-user-freedom/2013/01/21`. Accessed: 2020-12-05.

[6] Morten Warncke-Wang, Dan Cosley, and John Riedl. Tell me more: An actionable quality model for wikipedia. In *Proc. of Int. Symposium on Open Collaboration*, WikiSym '13, 2013.

[7] Andrew S Lan, Christopher G Brinton, Tsung-Yen Yang, and Mung Chiang. Behavior-based latent variable model for learner engagement. In *Proc. of Int. Conf. on Educational Data Mining*, 2017.

[8] M Akber Dewan, Mahbub Murshed, and Fuhua Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1, 2019.

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

[10] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. In *Proc. of the Twelfth Int. Conf. on Web and Social Media*, 2018.

[11] Marco Ponza, Paolo Ferragina, and Soumen Chakrabarti. On computing entity relatedness in wikipedia, with applications. *Knowledge-Based Systems*, 188, 2020.

[12] Jianwei Shi, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph 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.

[13] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. Towards an integrative educational recommender for lifelong learners. In *AAAI Conference on Artificial Intelligence*, 2020.

[14] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 1994.

[15] Radek Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2):1–19, 2015.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.

[16] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*, pages 171–180. Springer, 2013.

[17] Chen Lin and Min Chi. Intervention-bkt: Incorporating instructional interventions into bayesian knowledge tracing. In Alessandro Micarelli, John Stamper, and Kitty Panourgia, editors, *Proc. of Int. Conf. on Intelligent Tutoring Systems*, 2016.

[18] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor. D1.3 – initial content representations. `https://www.x5gon.org/wp-content/uploads/2019/10/X5GON_Deliverable_D1_3.pdf`. Accessed in: 2020-12-02.

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

[20] Radek Pelánek. Managing items and knowledge components: domain modeling in practice. *Educational Technology Research and Development*, 68(1):529–550, 2020.

Copyright - This document has been produced under the EC Horizon2020 Grant Agreement H2020-ICT-2014 /H2020-ICT-2016-2-761758. This document and its contents remain the property of the beneficiaries of the X5GON Consortium.