

VLEngagement: A Dataset of Scientific Video Lectures for Evaluating Population-based Engagement

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

m.bulathwela@ucl.ac.uk

Centre for Artificial Intelligence, University College London
London, UK

ABSTRACT

With the emergence of e-learning and personalised education, the production and distribution of digital educational resources have boomed. Video lectures have now become one of the primary modalities to impart knowledge to masses in the current digital age. The rapid creation of video lecture content challenges the currently established human-centred moderation and quality assurance pipeline, demanding for more efficient, scalable and automatic solutions for managing learning resources. Although a few datasets related to engagement with educational videos exist, there is still an important need for data and research aimed at understanding learner engagement with scientific video lectures. This paper introduces VLEngagement, a novel dataset that consists of content-based and video-specific features extracted from publicly available scientific video lectures and several metrics related to user engagement. We introduce several novel tasks related to predicting and understanding context-agnostic engagement in video lectures, providing preliminary baselines. This is the largest and most diverse publicly available dataset to our knowledge that deals with such tasks. The extraction of Wikipedia topic-based features also allows associating more sophisticated Wikipedia based features to the dataset to improve the performance in these tasks. The dataset, helper tools and example code snippets are available publicly at <https://github.com/sahanbull/context-agnostic-engagement>.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; Information extraction; • **Applied computing** → **Interactive learning environments**.

KEYWORDS

datasets, open education, engagement, video lectures, entity linking

ACM Reference Format:

Sahan Bulathwela, María Pérez-Ortiz, Emine Yilmaz and John Shawe-Taylor. 2020. VLEngagement: A Dataset of Scientific Video Lectures for Evaluating Population-based Engagement. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

Conference'17, July 2017, Washington, DC, USA

© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Formal evaluations have shown that intelligent tutoring systems produce similar learning gains as one-on-one human tutoring, which has the potential to increase student performance to around the 98 percentile in a standard classroom [3, 15, 42]. Additionally, intelligent tutors could effectively reduce by one-third to one-half the time required for learning [42], increase effectiveness by 30% as compared to traditional instruction [20, 21, 42], reduce the need for training support personnel by about 70% and operating costs by about 92% and facilitate education in developing countries [30, 40]. Thus, the idea of building intelligent tutoring systems that provide online personalised education has gained a lot of traction in the recent years and will continue to do so.

With more learning resources being created every day, automatic, scalable tools for quality assurance become essential [12]. Large educational resource repositories need scalable tools to understand/ estimate the engagement potential of newly added materials before exposing it to the learner audience [14]. Thus, estimating context-agnostic (also named here population-based) engagement of materials and releasing related datasets becomes a critical part of quality assurance, recommendation and information retrieval.

This work presents VLEngagement, a novel dataset that covers over 4000 peer-reviewed scientific video lectures constructed from a popular OER repository, VideoLectures.NET. The dataset has been proven very useful in some of our previous work, spanning both applications related to personalised and population based educational recommender systems. We believe the dataset has incredible potential for building intelligent educational and scientific recommender systems, specially given that similar datasets are usually proprietary and not publicly available. The dataset provides an extensive set of textual and video-specific features extracted from the lecture transcripts, together with Wikipedia topics covered in the lecture (via entity linking) and user engagement labels for each lecture. The dataset covers a wide set of tasks that are of crucial importance to building intelligent tutors, scalable quality assurance and understanding the features involved in population-based engagement. While video retrieval and ranking are actively researched areas, this dataset allows adapting algorithms specifically for scientific video content which is novel and critical to building information retrieval systems in education and science.

The dataset is particularly suited to solve the cold-start problem found in educational recommender systems, both when i) new users join the system and we may not have enough information about their context so we may simply recommend population-based engaging lectures for a specific query topic and ii) new educational content is released, for which we may not have user engagement

data yet and thus an engagement predictive model would be necessary. To the best of our knowledge, this is the first dataset to tackle such a task in education/scientific recommendations. The aim of the dataset is not to replace personalised recommenders by building population-based models, but rather to enable mixing personalised approaches with a meaningful population baseline/prior to solve the common cold-start problem and, given the context of the learner, rank the suitable material by their engagement potential [9].

2 RELATED WORK

The learning analytics and educational data mining communities have developed novel algorithms to trace knowledge in learners [32, 39] and provide personalised recommendations for educational material [10, 31]. All these approaches are focused on capturing and exploiting the *context* of the learner. This is, among others, the knowledge state, interests, preferences and learning goals, all of which are crucial variables to develop an effective personalised tutoring system. However, in a landscape where new educational resources are created and circulated at a rapid scale (e.g. Open Educational Resources (OER) [38] and Massively Open Online Courses [34]), there is a big gap of knowledge in our understanding of the features involved in *context-agnostic* (i.e. population-based) engagement of educational resources and the relationship between different measures of engagement, popularity and subjective assessment in general. Our work thus focuses on connecting content analytics to population-based learning analytics.

Learner engagement is a necessary prerequisite for acquiring knowledge from educational resources. Carini et al. has shown that student engagement positively correlates with desired learning outcomes such as critical thinking and better grades in a conventional classroom setting. Various studies have demonstrated that learner engagement plays a key role in successful achievement of expected learning outcomes in an online learning setting [34, 36]. *Engagement* is a loaded concept that can have different definitions to different communities. For example, engagement is measured using different metrics depending on the modality of the educational resource.

Most work related to modelling educational engagement attempts to model engagement as a function of the context of the learner [1, 4, 25]. Our work, on the other hand, proposes to model context-agnostic engagement through several content-based features of the educational resource. Context-agnostic engagement has been previously studied for video lectures, albeit from a more qualitative perspective, with general recommendations such as keeping videos short [23] and using conversational language for lecture delivery [5]. These recommendations help authors to create better educational videos, but none of these works address the need for predicting automatically highly engaging educational resources, which is crucial for retrieving and recommending educational material at scale.

2.1 Related Datasets

The interest in identifying engaging information goes beyond the educational domain and is investigated in numerous other fields. These works show that numerous feature verticals associated to content, such as *understandability*, *freshness*, *topic coverage*, *presentation* and *authority* exist [11]. Engagement (specifically watch

time) has been used as the main measure used for YouTube recommendations [16] and to predict engagement with general-purpose videos [43]. This is usually the case for most media recommenders and several datasets are available for such task.

Looking beyond videos, Wikipedia uses a review system to evaluate the quality of its articles and several attempts have been made to build machine learning predictive models using features such as text style, readability, structure, network, recency and review information [18, 41]. This Wikipedia article quality dataset [18] is publicly available although user engagement data is not included. Only explicit quality labels are provided. Similar datasets are available for automated essay scoring [37].

There are only a handful of publicly available datasets that are related to predicting engagement in videos. Large-scale datasets focused on predicting engagement with general purpose videos (such as the one in [43] that analyses engagement in YouTube) are common but these lack focus on educational material. Some of the features used by these works share some similarity to the ones used in this paper (such as video duration, category, language and topic features). However, a large part of the features are focused on the reputation of the YouTube channel. No textual features relating to understandability and presentation are used, since this may be of less importance for general purpose videos than for education.

The most relevant dataset in the literature to understanding engagement in video lectures, i.e. the work based on approximately 800 videos from a Massively Open Online Course (MOOC) [23], is not publicly available. In this case, the data was manually processed and authors provided a qualitative analysis of engagement, with some 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. This dataset is also not publicly available and does not focus on learner engagement or subjective assessment metrics.

Although online learning platforms such as EdX [23, 35], Khan Academy [28] and other platforms harvest valuable learner behavioural data that is created in an "in-the-wild" setting, the datasets are often not publicly released due to the proprietary nature of the content and the user data. This work addresses this significant gap of data by constructing and releasing a dataset with over 4,000 scientific video lectures (OERs) associated with explicit star ratings and implicit engagement signals from hundreds of thousands of informal learners consuming video lectures in an in-the-wild setting.

Finally, a different line of work focuses on studying and identifying engagement from the learners perspective, through the recording of user learning sessions and brainwaves [44] and the use of computer vision and affect recognition [26] to propose *automatic*, *semi-automatic* and *manual* techniques. Although multiple datasets exist to address this task, these datasets are usually collected in lab setting using a limited number of participants [19]. However, the focus of these datasets is to detect learner engagement using a set of multi-modal data related to the learner (brain waves, visual information, learning logs, etc.) rather than the features of the content itself. There are also a handful of recent public

datasets/competitions relating to how students interact with learning problems (e.g. assistments¹ or multiple choice questions²), but these datasets do not focus on engagement.

3 VLENGAGEMENT DATASET

The VLEngagement dataset is constructed using the aggregated video lectures consumption data coming from a popular scientific OER repository, VideoLectures.Net³. 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. It is noteworthy that the dataset consists of *scientific video lectures* that explain novel scientific work geared more towards postgraduate, PhD level learners and the scientific research community. Therefore, the learner audience of the video lectures in this dataset may significantly differ from one of a conventional MOOC platform.

The dataset provides a set of statistics aimed at studying population based engagement in video lectures, together with other conventional metrics in subjective assessment such as average star ratings and number of views. We believe the dataset will serve the community applying AI in Education to further understand what are the features of educational material that makes it engaging for learners.

3.1 Feature Extraction

The dataset provides three types of features as outlined in Table 2: i) content-based textual features, ii) Wikipedia entity linking features and iii) video-based features. Although our dataset is composed of video lectures data, the majority of our features (with exception of some of the features in the video-based category) can be used across different modalities of educational material (e.g. books) as they are computed only considering the text transcription. The transcriptions for the English lectures and the English translations of the non-English lectures are provided by the TransLectures project⁴.

In this section, we define how different features are calculated from the lecture transcription. These features have been identified from the related work and are categorised under different verticals of quality assurance in text articles [2, 18, 29, 41] and engagement with video lectures [23]. The verticals are for example understandability, topic coverage, presentation, freshness and authority [11]. The code for computing some of these features is available together with the dataset.

3.1.1 Content-based Features. For explaining the features based on content transcripts, several functions need to be introduced: i) $\text{count}(s)$ is a function that returns the number of tokens in string s , ii) $\text{count}(t, s)$ is a function that returns the number of occurrences of tokens in token set t in string s and iii) $\text{u_count}(t, s)$ returns the frequency of unique tokens from token set t in string s . String s can be the transcript text s_{tr} or the lecture title s_{title} . *Stop-word Presence Rate* and *Stop-word Coverage Rate* are calculated using Eq. 5 and 6 based on the work of Ntoulas et al. Textual features

defined by Eq. 7 through Eq. 12 are based on the work of Dalip et al. All definitions used the token sets provided in Table 6. More specifically, the content-based features extracted are the following:

- *Word Count* of lecture transcript s_{tr} :

$$\text{Word Count} = \text{count}(s_{tr}) \quad (1)$$

- *Title Word Count* of lecture s_{title} :

$$\text{Title Word Count} = \text{count}(s_{title}) \quad (2)$$

- *Document Entropy*, based on the work of Bendersky et al., is calculated over every word w in transcript s_{tr} as:

$$\text{Document Entropy} = \sum_{w \in s_{tr}} p_{s_{tr}}(w) \log p_{s_{tr}}(w), \quad (3)$$

$$\text{where } p_{s_{tr}}(w_i) = \frac{\text{count}(w_i, s_{tr})}{\text{Word Count}}.$$

- *FK Easiness* is computed using textatistic [24] for transcript s_{tr} using:

$$\text{FK Easiness} = 206.835 - 1.015 \left(\frac{\text{Word Count}}{\text{sen_count}(s_{tr})} \right) - 84.6 \left(\frac{\text{syll_count}(s_{tr})}{\text{Word Count}} \right) \quad (4)$$

where $\text{sen_count}(s_{tr})$ and $\text{syll_count}(s_{tr})$ returns the number of sentences and syllables in transcript s_{tr} respectively. FK Easiness proxies complexity of the language used giving a low score for complex language and vice versa.

- *Stop-word Presence Rate* of lecture transcript s_{tr} :

$$\text{Stop-word Presence Rate} = \frac{\text{count}(sw, s_{tr})}{\text{Word Count}} \quad (5)$$

- *Stop-word Coverage Rate* of lecture transcript s_{tr} :

$$\text{Stop-word Coverage Rate} = \frac{\text{u_count}(sw, s_{tr})}{\text{count}(sw)} \quad (6)$$

- *Preposition Rate* of the lecture transcript s_{tr} :

$$\text{Preposition Rate} = \frac{\text{count}(prep, s_{tr})}{\text{Word Count}} \quad (7)$$

- *Auxiliary Rate* of the lecture transcript s_{tr} :

$$\text{Preposition Rate} = \frac{\text{count}(auxi, s_{tr})}{\text{Word Count}} \quad (8)$$

- *To Be Rate* of lecture transcript s_{tr} :

$$\text{To Be Rate} = \frac{\text{count}(tobe, s_{tr})}{\text{Word Count}} \quad (9)$$

- *Conjunction Rate* of lecture transcript s_{tr} :

$$\text{Conjunction Rate} = \frac{\text{count}(conj, s_{tr})}{\text{Word Count}} \quad (10)$$

- *Normalisation Rate* of lecture transcript s_{tr} :

$$\text{Normalisation Rate} = \frac{\text{count}(norm, s_{tr})}{\text{Word Count}} \quad (11)$$

- *Pronoun Rate* of lecture transcript s_{tr} :

$$\text{Pronoun Rate} = \frac{\text{count}(pron, s_{tr})}{\text{Word Count}} \quad (12)$$

- *Published Date* of video lecture ℓ calculates the epoch time of publication date of the lecture in days [8]:

$$\text{Published Date} = \text{days}(\ell_{pub_date} - 1970/01/01) \quad (13)$$

¹<https://sites.google.com/site/assistmentsdata/home/assistent-2009-2010-data>

²<https://www.microsoft.com/en-us/research/event/diagnostic-questions-neurips2020/>

³www.videolectures.net

⁴www.translectures.eu

Various prior works provide the rationale behind the suitability of these features [11, 17, 23].

3.1.2 Wikipedia-based Features. The Wikipedia topics most connected to the lectures are identified using Wikification [6], an entity linking approach. Using the identified Wiki topics, four different feature groups are introduced with the dataset. They fall under the *Authority* and *Topic Coverage* verticals.

The *top-5 authoritative topic URLs* and *top-5 PageRank scores* features represent the *Topic Authority* feature vertical. Figure 1 (left) shows the summary of Wikipedia topics that are most authoritative (top 1 topic) in the lectures found in the dataset. When PageRank score [7] is computed, Wikipedia topics heavily connected to other topics (i.e. more semantically related) within the lecture will emerge. Hence, the top-ranking topics are the more authoritative topics within the context of topics in the lecture. During Wikification [6], a semantic graph is constructed where semantic relatedness ($SR(c, c')$) between each Wikipedia topic pair c and c' in the graph are calculated using:

$$SR(c, c') = \frac{\log(\max(|L_c|, |L_{c'}|) - \log(|L_c \cap L_{c'}|))}{\log |W| - \log(\min(|L_c|, |L_{c'}|))} \quad (14)$$

where L_c represents the set of topics with inwards links to Wikipedia topic c , $|\cdot|$ represents the cardinality of the set and W represents the set of all Wikipedia topics. This semantic relatedness graph is used for computing PageRank scores. It is noteworthy that "authority" of a learning resource entails author, organisation and content authority [11]. These features represent content authority. The top 5 topic URLs and their relative PageRank Score are included as two feature groups providing 10 distinct features for each video lecture.

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 c is calculated using:

$$\cos(s_{tr}, c) = \frac{TFIDF(s_{tr}) \cdot TFIDF(c)}{\|TFIDF(s_{tr})\| \times \|TFIDF(c)\|} \quad (15)$$

where $TFIDF(s)$ returns the TF-IDF vector of string s . Topics in the lecture are then ranked using this score. Figure 1 (right) shows the summary of Wikipedia Topics that are most covered (top 1 topic) in the lectures found in the dataset. The top 5 covered topic URLs and their cosine similarity scores are included as two additional feature groups providing 10 distinct features.

Topic authority and topic coverage features represent two different aspects of the content of a video lecture. Authoritative topics are the ones highly connected and dominant within the range of topics that are discussed in the lecture. An authoritative topic needs to have high semantic relatedness to other topics in the lecture. On the contrary, covered topics represent the heavy overlap between individual Wikipedia topics and the lecture transcript. Figure 1 gives further evidence of how these two feature groups are different from each other. The most emerging Wikipedia topics that are authoritative (left) in the lecture dataset are very different from the covered topics (right). The figure also shows that the authoritative topics are narrowly focused concepts (e.g. Machine Learning, Algorithm, Ontology, etc.) whereas the most covered topics tend to be more general topics (e.g. Time, Scientific Method, Unit, etc.).

Table 1: 14 types of lectures in the VLEngagement dataset and their abbreviation (Abbr.) and frequency (Freq).

Abbr.	Description	Freq.	Abbr.	Description	Freq.
vbp	Best Paper	16	vdb	Debate	30
vdm	Demonstration	124	viv	Interview	52
vid	Introduction	15	vit	Invited Talk	300
vkn	Keynote	115	vl	Lecture	2956
vop	Opening	31	oth	Other	15
vpa	Panel	44	vps	Poster	56
vpr	Promotional Video	23	vtt	Tutorial	269

3.1.3 Video-specific Features. A set of easily automatable features that are video specific are also included in the VLEngagement dataset. Features *Lecture Duration*, *In Chunks*, *Lecture Type* and *Speaker Speed* are calculated based on prior work [23]. *Lecture Duration* feature reports the duration of the video in seconds. *In Chunks* is a binary feature which reports *True* if the lecture consists of multiple videos, and *False* otherwise. *Lecture type* value is derived from the metadata. The possible values for this feature are described in Table 1.

A novel feature *Silence Period Rate (SPR)* is introduced using the "silence" tags that are present in the video lecture transcript. The feature is defined as:

$$SPR(\ell) = \frac{1}{D(\ell)} \sum_{t \in T(\ell)} D(t) \cdot I(N(t) = \text{"silence"}) \quad (16)$$

where t is a tag in the collection of tags $T(\ell)$ that belong to lecture ℓ , N returns the type of tag t and D returns the duration of tag t or lecture ℓ and $I(\cdot)$ is the indicator function (returning 1 when the condition is verified, 0 otherwise).

3.2 Labels

There are several target labels available in the VLEngagement dataset. These target labels are created by aggregating available explicit and implicit feedback measures in the repository. Mainly, the labels can be constructed as three different types of quantification's of learner subjective assessment of a video lecture. The relationship between these different subjective assessments metrics can be investigated with the VLEngagement dataset.

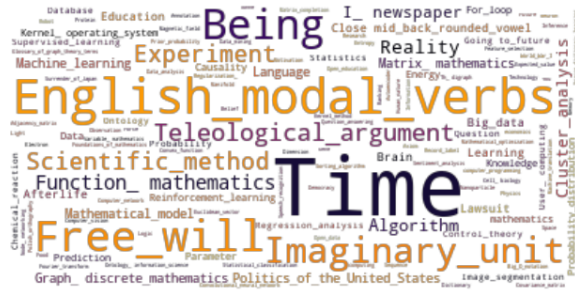
3.2.1 Explicit Rating. In terms of rating labels, *Mean Star Rating* is provided for the video lecture using a star rating scale from 1 to 5 stars. As expected, explicit ratings are scarce and thus only populated in a subset of resources (1250 lectures). Lecture records are labelled with -1 where star rating labels are missing. The data source does not provide access to ratings from individual users. Instead, only the aggregated average rating is available.

3.2.2 Popularity. A popularity-based target label is created by extracting the *View Count* of the lectures. The total number of views for each video lecture as of February 17, 2018 is extracted from the metadata and provided with the dataset.

3.2.3 Watch Time/Engagement. The majority of learner engagement labels in the VLEngagement dataset are based on watch time.



Most Authoritative Topic (PageRank – based)



Most Covered Topic (Cosine – based)

Figure 1: WordClouds summarising the distribution of the most authoritative (left) and most covered (right) Wikipedia topics in the dataset. Note that Computer Science and Data Science are the two dominant knowledge areas in our dataset.

Table 2: Features extracted and available in the VLEngagement dataset with their variable type (Continuous vs. Categorical) and their quality vertical.

Type	Feature	Quality Vertical
<i>Metadata features</i>		
cat.	Language (English, non-English)	—
cat.	Domain (STEM, Miscellaneous)	—
<i>Content-based features</i>		
con.	Word Count	Topic Coverage
con.	Title Word Count	Topic Coverage
con.	Document Entropy	Topic Coverage
con.	Easiness (FK Easiness)	Understandability
con.	Stop-word Presence Rate	Understandability
con.	Stop-word Coverage Rate	Understandability
con.	Preposition Rate	Presentation
con.	Auxiliary Rate	Presentation
con.	To Be Rate	Presentation
con.	Conjunction Rate	Presentation
con.	Normalisation Rate	Presentation
con.	Pronoun Rate	Presentation
con.	Published Date	Freshness
<i>Wikipedia-based features</i>		
cat.	Top-5 Authoritative Topic URLs	Authority
con.	Top-5 PageRank Scores	Authority
cat.	Top-5 Covered Topic URLs	Topic Coverage
con.	Top-5 Cosine Similarities	Topic Coverage
<i>Video-based features</i>		
con.	Lecture Duration	Topic Coverage
cat.	Is Chunked	Presentation
cat.	Lecture Type	Presentation
con.	Speaker speed	Presentation
con.	Silence Period Rate (SPR)	Presentation

We aggregate the user view logs and use the Normalised Engagement Time (NET) to compute the **Median of Normalised Engagement (MNET)**, as it has been proposed as the gold standard for engagement with educational materials in previous work [23]. We also calculate the **Average of Normalised Engagement (ANET)**. To have the MNET and ANET labels in the range [0, 1], we set the upper bound to 1 and derive Saturated MNET (SMNET) and Saturated ANET (SANET) respectively. Final SMNET (*Median Engagement*) for

Table 3: Labels included in the VLEngagement dataset with their variable type, value interval and category.

Type	Label	Interval	Category
cont.	Mean Star Rating	[1, 5]	Explicit Rating
cont.	View Count	(5, ∞)	Popularity
cont.	SMNET (Eq. 17)	(0, 1)	Watch Time
cont.	SANET (Eq. 18)	[0, 1]	Watch Time
cont.	Std. of NET	(0, 1)	Watch Time
cont.	Number of User Sessions	(5, ∞)	Watch Time
cont.	Engagement Times (NET)	[0, 1]	Watch Time

lecture ℓ is computed as:

$$SMNET(\ell) = \max(MNET(\ell), 1) \tag{17}$$

Similarly, *Average Engagement* is calculated using:

$$SANET(\ell) = \max(ANET(\ell), 1). \tag{18}$$

The standard deviation of NET for each lecture (*Std of Engagement*) is reported, together with the *Number of User Sessions* used for calculating MNET. These additional features allow future studies to incorporate the degree of uncertainty and statistical confidence in the engagement labels (e.g. in their loss functions or performance metrics). Furthermore, the individual NET values for each lecture are also provided with the dataset. This allows having much more insight into the true distribution of NET for individual lectures rather than summary statistics. This data will allow future studies to refine engagement labels or use more sophisticated methods to predict engagement.

3.3 Anonymity

We restrict the final dataset to lectures (ℓ) that have been viewed by at least 5 unique users to have reliable engagement measurements. Additionally, a regime of techniques are used for preserving the anonymity of the lectures in order to preserve the identities of the authors/lecturers. The motivation behind this decision is to avoid authors of the video lectures having unanticipated effects on their reputation by associating implicit learner engagement values to their content.

Rarely occurring values in *Lecture Type* feature were grouped together to create the *other* category found in Table 1. *Language*

feature is grouped into en and non-en categories. Similarly, Domain category groups Life Sciences, Physics, Technology, Mathematics, Computer Science, Data Science and Computers subjects to stem category and the other subjects to misc category. Rounding is used with *Published Date*, rounding to the nearest 10 days. *Lecture Duration* is rounded to the nearest 10 seconds. Gaussian white noise (10%) is added to *Title Word Count* feature and rounded to the nearest integer.

3.4 Final Dataset

The final dataset includes lectures that are published between September 1, 1999 and October 1, 2017. The engagement labels are created from 155,850 user views logged between December 8, 2016 and February 17, 2018. The final dataset consists of 4,046 lectures across 21 subjects (eg. Computer Science, Philosophy, etc.) that are categorised into STEM and Miscellaneous domains. The dataset, helper tools and example code snippets are available publicly⁵.

4 SUPPORTED TASKS

This section introduces the reader to the tasks that the dataset could be used for. The main application areas of these tasks are quality assurance in open education and scientific content recommenders and understanding and predicting population engagement in an online learning setting. Tasks 1 and 2 are demonstrated in this paper. Tasks 3-6 have been partially tackled in our prior work [8]. Tasks 7-8 are novel.

We establish two main tasks, which we mainly focus on in this paper, that can be objectively addressed using the VLEngagement dataset using a supervised learning approach. These are:

- (1) **Task 1: Predicting context-agnostic (population-based) engagement of video lectures:** The dataset provides a set of relevant features and labels to construct machine learning models to predict context-agnostic engagement in video lectures. The task can be treated as a regression problem to predict the different engagement labels.
- (2) **Task 2: Ranking of video lectures based on engagement:** Building predictive models that could rank lectures based on their context-agnostic engagement could be useful in the setting of an educational recommendation system, including tackling the cold-start problem associated to new video lectures. The task can be treated as a ranking problem to predict the global/relative ranking of video lectures.

We further identify several auxiliary tasks that can also be addressed with this dataset:

- **Task 3: Features influencing engagement:** Uncovering the role of different textual and video-specific features involved in several statistics of population-based engagement.
- **Task 4: Influence of topics in engagement:** Understand the role that the topical content in the lecture play on population based engagement (with link to the Wikipedia pages of these topics).
- **Task 5: Disentangle different factors from engagement:** Compare features involved in engagement for different video

lecture types, language and knowledge areas (e.g. STEM vs non-STEM lectures).

- **Task 6: Comparing different measures of implicit and explicit subjective assessment:** Analyse the differences between engagement vs mean star ratings and number of views to identify the strengths and weaknesses of the different feedback types.
- **Task 7: Unsupervised learning to understand the distribution of video lectures:** Cluster video lectures according to the provided features to understand their distribution. Identification of formal patterns that depict similarities and differences between lectures could be insightful.
- **Task 8: Deducing the structure of knowledge:** The co-occurrence patterns of topics within the video lectures provide a great source of data to understand inter-topic relationships and how knowledge is structured. Work in this direction can be used in identifying related materials and accounting for novelty in educational recommendation [10].
- **Task 9: Contrasting to other educational datasets:** The lectures in the VLEngagement dataset are scientific videos, thus it may be meaningful to study if similar patterns for engagement hold across other educational datasets that come from other settings (e.g.: MOOCs).

We propose two baseline models addressing the main tasks (1 and 2) in section 5.

4.1 Evaluating Performance

We identify *Root Mean Squared Error (RMSE)* as a suitable metric for 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)* and *Pairwise Ranking Accuracy (Pairwise)*. SROCC is suitable for comparing between ranking models that create global rankings (e.g. point-wise ranking algorithms). However, pairwise ranking accuracy is more intuitive for this task as it represents the fraction of pairwise comparisons where the model could predict the more engaging lecture. There is more than one unique solution for this problem, especially when there is error associated with the ranking model [22].

We use 5-fold cross validation to evaluate model performance with tasks 1 and 2. We release the folds together with the dataset, to allow for fair comparisons to the baselines. The five folds can be identified using the *fold* column in the dataset. 5-fold cross validation also allows reporting the *standard error* ($1.96 \times \text{Standard Deviation}$) of the performance estimate, which we include in our results in tables 4 and 5.

5 EXPERIMENTS AND BASELINES

Prior work on similar tasks identify ensemble models [8, 41] to be the best performing models with the main tasks described in section 4. We use *Random Forests Regressor (RF)* and *Gradient Boosting Machines (GBM)* for constructing baselines. We use *SMNET* labels as the target variable for both engagement prediction and video lecture ranking tasks. No pre-processing or cleaning steps are necessary.

⁵<https://github.com/sahanbull/context-agnostic-engagement>

5.1 Features and Labels for Baseline Models

All the features outlined in the content-based and video-based sections in Table 2 are included in the baseline models. However, due to the large amount of topics available in the Wikipedia-based feature groups, we restrict the feature set by adding only the *most authoritative topic URL* and *most covered topic URL*, where both the features are added to the baseline models as categorical variables. Practitioners are encouraged to try further encodings of these variables, as it will likely have a great impact in the performance.

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.

This allows identifying the performance gain achieved through adding each new group of features.

Our 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.

5.2 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-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.

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

Feature Set	RMSE	
	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).

Feature Set	GBM		RF	
	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

5.3 Limitations and Opportunities

This dataset has several limitations that are noteworthy. For example, as the topics in Figure 1 indicate, this dataset is dominated with Computer Science and Data Science related lectures that are mainly delivered in English. In addition, the majority of lectures in the dataset are research talks, narrowing down the style and type of data. These limitations cast significant uncertainty regarding the generalisation of the prediction models to more diverse types of educational video lectures. Although VLEngagement dataset is large compared to the rest of educational engagement datasets available, it still suffers from a limitation in the variety of its data.

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 [27]. Analysing facial expressions and affective states is another alternative approach to representing engagement [19]. 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 [23, 43], 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.

6 CONCLUSIONS AND FUTURE DIRECTIONS

Identifying the need for understanding context-agnostic engagement prediction to 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

Table 6: Tokens used for Feature Extraction.

Token Set	Description	Tokens
sw	Stopwords	all, show, anyway, fifty, four, go, mill, find, seemed, one, whose, re, herself, whoever, behind, should, to, only, under, herein, do, his, get, very, de, none, cannot, every, during, him, did, cry, beforehand, these, she, thereupon, where, ten, eleven, namely, besides, are, further, sincere, even, what, please, yet, couldn't, enough, above, between, neither, ever, across, thin, we, full, never, however, here, others, hers, along, fifteen, both, last, many, whereafter, wherever, against, etc, s, became, whole, otherwise, among, via, co, afterwards, seems, whatever, alone, moreover, throughout, from, would, two, been, next, few, much, call, therefore, interest, themselves, thr, until, empty, more, fire, latterly, hereby, else, everywhere, former, those, must, me, myself, this, bill, will, while, anywhere, nine, can, of, my, whenever, give, almost, is, thus, it, cant, itself, something, in, ie, if, inc, perhaps, six, amount, same, wherein, beside, how, several, whereas, see, may, after, upon, hereupon, such, a, off, whereby, third, i, well, rather, without, so, the, con, yours, just, less, being, indeed, over, move, front, already, through, yourselves, still, its, before, thence, somewhere, had, except, ours, has, might, thereafter, then, them, someone, around, thereby, five, they, not, now, nor, name, always, whither, t, each, become, side, therein, twelve, because, often, doing, eg, some, back, our, beyond, ourselves, out, for, bottom, since, forty, per, everything, does, three, either, be, amongst, whereupon, nowhere, although, found, sixty, anyhow, by, on, about, anything, theirs, could, put, keep, whence, due, ltd, hence, onto, or, first, own, seeming, formerly, into, within, yourself, down, everyone, done, another, thick, your, her, whom, twenty, top, there, system, least, anyone, their, too, hundred, was, himself, elsewhere, mostly, that, becoming, nobody, but, somehow, part, with, than, he, made, whether, up, us, nevertheless, below, un, were, toward, and, describe, am, mine, an, meanwhile, as, sometime, at, have, seem, any, fill, again, hasn't, no, latter, when, detail, also, other, take, which, becomes, yo, towards, though, who, most, eight, amongst, nothing, why, don, noone, sometimes, together, serious, having, once, hereafter
conj	Conjunctions	and, but, or, yet, nor
norm	Normalizations	-tion, -ment, -ence, -ance
tobe	To-be Verbs	be, being, was, were, been, are, is
prep	Prepositions	aboard, about, above, according to, across from, after, against, alongside, alongside of, along with, amid, among, apart from, around, aside from, at, away from, back of, because of, before, behind, below, beneath, beside, besides, between, beyond, but, by means of, concerning, considering, despite, down, down from, during, except, except for, excepting for, from among, from between, from under, in addition to, in behalf of, in front of, in place of, in regard to, inside of, inside, in spite of, instead of, into, like, near to, off, on account of, on behalf of, onto, on top of, on, opposite, out of, out, outside, outside of, over to, over, owing to, past, prior to, regarding, round about, round, since, subsequent to, together, with, throughout, through, till, toward, under, underneath, until, unto, up, up to, upon, with, within, without, across, long, by, of, in, to, near, of, from
auxi	Auxiliary Verbs	will, shall, cannot, may, need to, would, should, could, might, must, ought, ought to, can't, can
pron	Pronouns	i, me, we, us, you, he, him, she, her, it, they, them, thou, thee, ye, myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself, my, mine, his, hers, yours, ours, theirs, its, our, that, their, these, this, those

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 [33]. A wider range of features that capture textual, audio-visual and presentation slides related patterns will be constructed [35]. 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 5.3, 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 [9].

ACKNOWLEDGMENTS

This research is part of the EU's Horizon 2020 grant No 761758 (www.x5gon.org) and partially funded by the EPSRC Fellowship titled "Task Based Information Retrieval", under grant No EP/P024289/1.

REFERENCES

- [1] Carole R Beal, Lei Qu, and Hyokyong Lee. 2006. Classifying learner engagement through integration of multiple data sources. In *Proc. of AAAI Conference on Artificial Intelligence*.
- [2] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-biased Ranking of Web Documents. In *Proc. of ACM Int. Conf. on Web Search and Data Mining*.
- [3] Benjamin S. Bloom. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13, 6 (1984), 4–16. <http://www.jstor.org/stable/1175554>
- [4] Fernanda Bonafini, Chungil Chae, Eunsung Park, and Kathryn Jablckow. 2017. How much does student engagement with videos and forums in a MOOC affect their achievement? *Online Learning Journal* 21, 4 (2017).
- [5] Cynthia J Brame. 2016. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education* 15, 4 (2016).
- [6] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating Documents with Relevant Wikipedia Concepts. In *Proc. of Slovenian KDD Conf. on Data Mining and Data Warehouses (SiKDD)* (Ljubljana, Slovenia).
- [7] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of Int. Conf. on World Wide Web*.
- [8] Sahan Bulathwela, Maria Perez-Ortiz, Aldo Lipani, Emine Yilmaz, and John Shawe-Taylor. 2020. Predicting Engagement in Video Lectures. In *Proc. of Int. Conf. on Educational Data Mining (EDM '20)*. <https://arxiv.org/pdf/2006.00592.pdf>
- [9] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. 2020. Towards an Integrative Educational Recommender for Lifelong Learners. In *AAAI Conference on Artificial Intelligence (AAAI '20)*.
- [10] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. 2020. TrueLearn: A Family of Bayesian Algorithms to Match Lifelong Learners to Open Educational Resources. In *AAAI Conference on Artificial Intelligence (AAAI '20)*.
- [11] Sahan Bulathwela, Emine Yilmaz, and John Shawe-Taylor. 2019. Towards Automatic, Scalable Quality Assurance in Open Education. https://www.k4all.org/wp-content/uploads/2019/08/IJCAI_paper_on_quality.pdf. In *Workshop on AI and the United Nations SDGs at Int. Joint Conf. on Artificial Intelligence*.
- [12] Anthony F. Camilleri, Ulf Daniel Ehlers, and Jan Pawlowski. 2014. *State of the art review of quality issues related to open educational resources (OER)*. Publications Office of the European Union 2014, Vol. 52 S. - JRC Scientific and Policy Reports.
- [13] Robert M Carini, George D Kuh, and Stephen P Klein. 2006. Student engagement and student learning: Testing the linkages. *Research in higher education* 47, 1 (2006).
- [14] K.I. Clements and J.M. Pawlowski. 2012. User-oriented quality for OER: understanding teachers' views on re-use, quality, and trust. *Journal of Computer Assisted Learning* 28, 1 (2012).
- [15] Albert Corbett. 2001. Cognitive Computer Tutors: Solving the Two-Sigma Problem. In *User Modeling 2001*, Mathias Bauer, Piotr J. Gmytrasiewicz, and Julita Vassileva (Eds.). Springer Berlin Heidelberg, 137–147.
- [16] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proc. of ACM Conf. on Recommender Systems*.
- [17] Daniel H. Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2017. 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).
- [18] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2011. Automatic Assessment of Document Quality in Web Collaborative Digital Libraries. *Journal of Data and Information Quality* 2, 3 (Dec. 2011).
- [19] M Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1.
- [20] JD Fletcher. 1988. Intelligent training systems in the military. *Defense applications of artificial intelligence: Progress and prospects*. Lexington, MA: Lexington Books (1988).
- [21] JD Fletcher. 1996. Does this stuff work? Some findings from applications of technology to education and training. In *Proceedings of conference on teacher education and the use of technology based learning systems*. Society for Applied Learning Technology Warrenton.
- [22] Johannes Fürnkranz and Eyke Hüllermeier. 2011. *Preference Learning and Ranking by Pairwise Comparison*. Springer Berlin Heidelberg, Berlin, Heidelberg, 65–82. https://doi.org/10.1007/978-3-642-14125-6_4
- [23] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In *Proc. of the First ACM Conf. on Learning @ Scale*.
- [24] E. Hengel. 2017. *Publishing while Female. Are women held to higher standards? Evidence from peer review*. Cambridge Working Papers in Economics 1753.
- [25] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, and Syed Muhammad Raza Abidi. 2018. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Computational Intelligence and Neuroscience* 2018, 6347186 (2018).
- [26] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. 2018. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 1–8.
- [27] Andrew S Lan, Christopher G Brinton, Tsung-Yen Yang, and Mung Chiang. 2017. Behavior-Based Latent Variable Model for Learner Engagement. In *Proc. of Int. Conf. on Educational Data Mining*.
- [28] Zachary MacHardy and Zachary A. Pardos. 2015. Evaluating The Relevance of Educational Videos using BKT and Big Data. In *Proc. of Int. Conf. on Educational Data Mining*.
- [29] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting Spam Web Pages Through Content Analysis. In *Proc. of Int. Conf. on World Wide Web*.
- [30] Benjamin D Nye. 2015. Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education* 25, 2 (2015), 177–203.
- [31] Zachary A. Pardos and Weijie Jiang. 2020. Designing for Serendipity in a University Course Recommendation System. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (Frankfurt, Germany) (LAK '20)*, 350–359.
- [32] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems* 28, 505–513.
- [33] Marco Ponza, Paolo Ferragina, and Soumen Chakrabarti. 2020. On Computing Entity Relatedness in Wikipedia, with Applications. *Knowledge-Based Systems* 188 (2020), 105051.
- [34] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. 2014. Learning latent engagement patterns of students in online courses. In *Proc. of AAAI Conference on Artificial Intelligence*.
- [35] Jianwei Shi, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph Ewerth. 2019. 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 (Nice, France) (SALMM '19)*. Association for Computing Machinery, New York, NY, USA, 11–19. <https://doi.org/10.1145/3347451.3356731>
- [36] Stefan Slater, Ryan Baker, Jaclynn Ocumpaugh, Paul Inventado, Peter Scupelli, and Neil Heffernan. 2016. Semantic Features of Math Problems: Relationships to Student Learning and Engagement. In *Proc. of Int. Conf. on Educational Data Mining*.
- [37] Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*.
- [38] UNESCO. 2019. Open Educational Resources (OER). <https://en.unesco.org/themes/building-knowledge-societies/oer>. Accessed: 2020-04-01.
- [39] Jill-Jënn Vie and Hisashi Kashima. 2019. Knowledge tracing machines: Factorization machines for knowledge tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 750–757.
- [40] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fusco Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11, 1 (2020), 1–10.
- [41] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell Me More: An Actionable Quality Model for Wikipedia. In *Proc. of Int. Symposium on Open Collaboration (WikiSym '13)*.
- [42] Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.
- [43] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Beyond Views: Measuring and Predicting Engagement in Online Videos. In *Proc. of the Twelfth Int. Conf. on Web and Social Media*.
- [44] Fangli Xu, Lingfei Wu, K. P. Thai, Carol Hsu, Wei Wang, and Richard Tong. 2019. MUTLA: A Large-Scale Dataset for Multimodal Teaching and Learning Analytics. *CoRR abs/1910.06078* (2019). arXiv:1910.06078 <http://arxiv.org/abs/1910.06078>