BiGBERT: Classifying Educational Web Resources for Kindergarten-12th Grades

Garrett Allen[1], Brody Downs[1], Aprajita Shukla[1], Casey Kennington[1], Jerry Alan Fails[1], Katherine Landau Wright[2], Maria Soledad Pera[1]

[1] Dept. of Computer Science - Boise State University - Boise, ID
[2] Dept. of Literacy, Language and Culture - Boise State University - Boise, ID
cast-group@boisestate.edu

Abstract. In this paper, we present BiGBERT, a deep learning model that simultaneously examines URLs and snippets from web resources to determine their alignment with children's educational standards. Preliminary results inferred from ablation studies and comparison with baselines and state-of-the-art counterparts, reveal that leveraging domain knowledge to learn domain-aligned contextual nuances from limited input data leads to improved identification of educational web resources.

Keywords: web classification, BERT, educational standards

1 Introduction

Web resource classification is a well-explored area in Information Retrieval [15]. Recently, the field has seen an influx of research related to domain-specific classification, especially within the legal, financial and medical domains [11, 18, 36]. Classification in the domain of *education*, however, remains relatively unexplored. As a broad term, education applies to a variety of classification tasks. Prior work includes classifying educational resources based on "the strength of the educative resource [as] a property evaluated cumulatively by the target audience of the resource (e.g., students or educational experts)" using a Support Vector Machine (SVM) [16]. This model, however, relies heavily on manually-annotated data and is applicable only to computer science education. Xia [32] also uses an SVM to classify resources supporting instruction, whereas EduBERT [7] detects college-level forum posts written by struggling students. In general, efforts in this area classify resources for unspecified age groups, adult students, limited subject areas, instructors or institutional-level insights. There is a gap in the literature regarding recognizing educational web resources for children ages 6-18 in grades Kindergarten-12 (**K-12**). Educational standards, such as the United States' Common Core State Standards (CCSS) and the Next Generation Science Standards (NGCS), provide learning outcomes for K-12 students. For example, a grade 1 learning outcome from CCSS states "Identify the main topic and retell key details of a text" [19]. We posit that domain knowledge obtained from these standards can inform the classification of children's educational web resources.

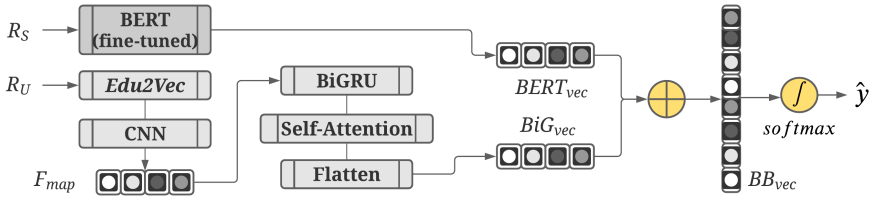

Fig. 1. BiGBERT architecture (R_U and R_S denote the URL and snippet,resp.).

Regardless of the domain, classifiers tend to rely on features inferred from HTML page content [9, 28]. Processing full web pages requires high computational power, large data storage, and time to retrieve [25] as web pages are often dynamic and contain pictures, videos, or scripts in addition to text [26]. To address some of these constraints, state-of-the-art approaches examine only URLs [14, 26]. Unfortunately, URLs are not always comprised of meaningful tokens (i.e., valid terms), which may cause misclassifications. Consider the URL *https://www.youtube.com/watch?v=pX3V9hoX1eM* for a YouTube video by National Geographic For Kids related to animals. In this case, meaningful tokens include "youtube" and "watch," neither of which indicates the corresponding resource is child-friendly.

Mindful of the aforementioned limitations, in this paper, we introduce **BiGBERT**, a Bidirectional Gated Recurrent Unit (BiGRU) with BERT that recognizes educational web resources for children. In particular, we focus on educational resources that inform on subjects for grades K-12, such as language arts, science and social studies, described in CCSS, NGCS, and Idaho Content Standards (ICS). As illustrated in Figure 1, BiGBERT has two main components: a URL and a snippet vectorizer. To vectorize URLs, we combine the domain-specific embeddings from *Edu2Vec* [3] with a BiGRU and a self attention layer. Shen et al. [27] show that using summaries instead of full page content results in comparable classification performance, thus we use snippets in place of full content. To vectorize snippets, we fine tune the transformer model BERT [8] using educational standards. Last, we concatenate the snippet and URL vectors and apply a softmax function to determine the class of a web resource.

With our work, we seek to answer these research questions: **RQ1**: Do URLs provide sufficient indication that resources are educational?; **RQ2**: Do snippets along with URLs help identify educational resources?; and **RQ3**: Does domain-specific knowledge affect identification of educational resources? Our main contribution is a hybrid strategy that simultaneously considers resource URL and snippet, while informing domain-dependent learning with minimal educational data for determining resource alignment to K-12 educational standards. We envision BiGBERT (https://github.com/BSU-CAST/BiGBERT) as groundwork to support other Information Retrieval tasks, e.g., easing access to online resources supporting K-12 curriculum-related information discovery tasks.

2 BiGBERT

In this section, we detail how BiGBERT simultaneously leverages features from the URL (R_U) and snippet (R_S) of a web resource R for classification purposes. BiGBERT is trained using a batch size of 128, binary cross-entropy loss function, and RMSProp optimizer [30] with momentum=0.2 and learning rate=0.001.

URL Vectorizer. BiGBERT tokenizes R_U into a sequence of terms T by splitting on non-alphanumeric symbols (e.g., periods, dashes and forward slashes) and using SymSpell [13] to perform word segmentation as URLs tend to compound words together (e.g., changing *stackoverflow* to *stack overflow*). Each token $t_i \in T$ is mapped to its corresponding word embedding. If t_i is not part of the embedding dictionary, we attribute this to a possible misspelling or spelling variation, and thus attempt a correction using a single edit distance operation (i.e., replacing, adding, or removing a character). If t_i is still not in the dictionary, we discard it to ensure only meaningful tokens remain.

To learn a representation of R_U, BiGBERT uses the *Edu2Vec* word embeddings dictionary [3] as it incorporates domain knowledge from NGCS, CCSS, and ICS. These standards serve as structured knowledge sources to identify terms, topics, and subjects for K-12 grades, enabling BiGBERT to emphasize K-12 curriculum concepts in R_U that may be overlooked by general-purpose pre-trained embeddings. Rather than analyzing independent embeddings, we design BiGBERT to scrutinize context-sensitive indications from T. Inspired by Rajalakshmi et al. [24] and in response to URLs not following traditional language syntax, we examine groups of embeddings (i.e., trigrams) using a Convolutional Neural Network (CNN)–a fast, effective, and compact method [20] to generate feature vectors from trigrams. The convolution results in a feature map F_{map}=<F_1,F_2,...,F_x>, $\forall_{f=1..x}\ F_f = relu(w.x_{i:i+m-1}+b_u)$, where the rectified linear function $relu$ is applied to the dot product of a kernel w with a window of embeddings $x_{i:i+m-1}$ in T of size m=3; b_u is a bias term. To explore long term dependencies of features that may appear far apart BiGBERT uses a BiGRU network, as it captures context information in a forwards and backwards direction. A self-attention layer then determines the importance of features identified by the CNN and BiGRU. This is followed by a flatten and dense layer that yields a single feature vector representation of R_U of size 128, denoted $\boldsymbol{BiG_{vec}}$.

Snippet Encoding. As snippets are a few sentences long, unlike URLs which are at most a few words, we require a model that can scrutinize each snippet as a whole. Hence, we incorporate the state-of-the-art transformer model BERT [8] into BiGBERT's design. BERT's ability to process sequences up to a maximum size of 512 tokens enables BiGBERT to exploit the sequential, contextual information within R_S in its entirety. Additionally, BERT's architecture consisting of 12 transformer blocks and self-attention heads ensures the learning of rich contextual information from each snippet. As such, we tokenize R_S into a sequence of sentences, encode it to BERT's specifications, and use BERT to attain an aggregate feature vector representation of size 768, denoted $\boldsymbol{BERT_{vec}}$.

On domain-dependent tasks like the one we address here, BERT benefits from fine-tuning [29]. Thus, we adjust traditional BERT to our definition of ed-

ucation by exploiting established educational standards. We perform fine-tuning as described in [29], training[3] BERT embeddings as an educational text classifier by adding a linear classification layer which uses binary cross entropy as loss and the Adam optimizer with learning rate=$1e^{-5}$.

Classification. To leverage evidence of educational alignment inferred from R_U and R_S, we concatenate BiG_{vec} with $BERT_{vec}$ as BB_{vec}. BiGBERT then invokes a fully connected layer on BB_{vec} that uses a softmax activation function to produce a probability distribution $\hat{\mathbf{y}}$ over each class, educational and not, such that $\hat{y} \in [0, 1]$. This function ensures that the sum of the probabilities per class adds up to one. The class predicted for R is the one with the highest probability.

3 Experiments & Discussion

We conducted empirical explorations to answer the research questions that guided our work. Below we discuss our experimental set up and results.

Set-up. There is no **dataset**[4] we can use to assess the proposed task. Thus, we build one using URLs (with text in English) from *Alexa Top Sites* [2]–based on the well-known Open Directory Project (ODP) [6, 22]. We treat as educational the 1,273 URLs in subcategories *Pre-School* and *School Time* from *Kids & Teens*. We also randomly select 3,998 non-educational URLs uniformly distributed among *Adult*, *Business*, *Recreation*, and *Games*. To validate that dataset labels align (or not) with our definition of educational, an education expert annotated a representative sample (n = 527). As in [23], we calculate the accuracy between the two annotations (Alexa vs. expert) per sample, obtaining an inter-annotator agreement of 94.7%. For performance assessment, we use **Accuracy**, a common classification metric, along with False Positive (**FPR**) and False Negative (**FNR**) ratios, to offer insights on the type of misclassified resources.

To the best of our knowledge, there are no domain-specific classifiers that we can use to contextualize BiGBERT's performance. Thus, we optimize and adapt several classifiers to detect K-12 web resources: (i) **BoW**[5] [14], a bag-of-words model that computes cosine similarity between a vectorized resource URL and ODP category descriptions to determine the resource's respective category (note that we use the text of learning outcomes from educational standards in lieu of category descriptions); (ii) **BGCNN** [26], a model based on a BiGRU with a CNN which identifies child-friendly URLs; (iii) **BERT4TC** [35], a text classifier that uses a BERT encoder to perform topic and sentiment classification, and (iv) **Hybrid-NB** [1], a hybrid model which examines both URL and content of websites to determine their target audience (i.e., Algerian users). Reported results for BGCNN and BERT4TC are the average of 5-fold cross validation. Additionally, we explore **variations** of BiGBERT where **U**, **S**, and **E** indicate when

[3] For fine-tuning we use 2,655 text passages from NGCS, CCSS, and ICS along with 2,725 from the Brown corpus [5, 12].

[4] Due to Terms of Use for Alexa Top Sites, we are unable to share this dataset.

[5] We explored SVM as an additional baseline, which performed similarly to BoW and is excluded for brevity.

Table 1. Experimental results. **U** and **S** applied to URL and snippet only; **E** augmented with educational data. * and † significant w.r.t. BiGBERT and non-educational counterpart, resp. Significance determined with McNemar's test, $p<0.05$.

Row	Type	Models	Accuracy	FPR	FNR
1	Baseline	BoW	.7205 *	.115	.796
2	State-of the-art	BGCNN	.8399 *	.073	.432
3		BERT4TC	.9353 *	.041	.140
4		Hybrid-NB	.8600 *	.145	.123
5	Ablation Study	BiGBERT-U	.8276 *	.073	.484
6		BiGBERT-U-E	.8287 * †	.072	.483
7		BiGBERT-S	.9374 *	.027	.175
8		BiGBERT-S-E	.9334 *	.038	.155
9		BiGBERT-U-S	.9381 *	.035	.146
10		BiGBERT	**.9533** †	**.027**	**.106**

BiGBERT examines only URLs, snippets, and infuses educational information, respectively. Finally, through an ablation study, we showcase the contributions of the URL and snippet vectorizers towards the overall architecture of BiGBERT.

Results and Discussion. We summarize our results in Table 1.

Do URLs provide sufficient indication that resources are educational? Reports in [26] showcase the effectiveness of only examining URLs to identify sites as child-friendly. This motivates us to study the applicability of the approach for detecting educational web resources targeting K-12 populations. The accuracy of BoW does not surpass the 75% mark attained via a naive baseline (one always predicting non-educational due to the unbalanced nature of our dataset). BGCNN, BiGBERT-U, and BiGBERT-U-E outperform more traditional models with accuracies in the low 80 percentile. We attribute the increase in performance to the fact that state-of-the-art models do not assume URL token independence, unlike BoW. Results from our analysis indicate that when semantic and context-rich information is available, URLs are a valuable source to inform classification. The number of misclassified educational resources in this case, however, is high as nearly half of educational samples, which comprise 25% of our data, are being labelled non-educational (see respective FNR). This leads us to investigate additional information sources that can contribute to the classification process.

Do snippets along with URLs help identify educational resources? As content analysis is a staple of classification, it is logical to consider knowledge inferred from snippets to better support the classification of K-12 educational web resources. This is demonstrated by significant performance improvements of Hybrid-NB, BiGBERT-U-S, and BiGBERT over counterparts solely looking at URLs (BoW and BGCNN). In fact, BiGBERT significantly outperforms hybrid models in accuracy and FPR. Fewer false positives means lower likelihood for potentially inappropriate sites being labelled educational, which is of special importance given the domain and audience of our work. The results suggest that snippets, combined with URLs, do help identify educational resources. How-

ever, the higher FNR of BiGBERT-U-S compared to Hybrid-NB, again points to the misclassification of educational resources. This can be seen on samples like *www.sesamestreet.org*, recognized as educational by Hybrid-NB but overlooked by BiGBERT-U-S. This would suggest that the lack of explicit domain knowledge is a detriment to BiGBERT-U-S.

Does domain-specific knowledge affect identification of educational resources? BiGBERT's accuracy increases when using Edu2Vec and fine-tuned BERT embeddings (rows 9 vs 10 in Table 1). To determine whether the improvement is the result of explicitly infusing educational knowledge into the classification process, we compare BiGBERT-U and BiGBERT-S with educationally-augmented counterparts. Our experiments reveal a significant decrease in FPR and FNR between BiGBERT-U and BiGBERT-U-E; non significant between BiGBERT-S and BiGBERT-S-E. Unlike for URL variations, BiGBERT-S-E's performance improved only in FNR after augmentation. We attribute this to the relatively small training set used for fine-tuning in comparison to the initial pre-training set for BERT, leading to less new contextual information learned by the standard transformer model. Nonetheless, the significant increases in accuracy and decreases in FPR and FNR for BiGBERT when compared to BiGBERT-U-S suggest that domain-specific knowledge can have a positive effect on the classification of educational resources. This is illustrated by the URL *www.xpmath.com*, a site to support math education in grades 2-9, that is labelled non-educational by BiGBERT-U-S, yet it is correctly recognized as educational by BiGBERT.

4 Conclusion and Future Work

In this paper, we focused on a relatively unexplored area: identification of educational web resources for K-12 populations. We introduced BiGBERT based on a hybrid, deep learning architecture that relies on contextual analysis strategies alongside educational knowledge sources to capture features that best showcase resource alignment with K-12 subjects. Results from our experiments demonstrate that classifiers of educational K-12 web resources benefit from concurrently accounting for snippets and URLs. Further, via an ablation study we validate BiGBERT's design; specifically the need for the infusion of educational domain knowledge. Outcomes from our work align with [21], regarding leveraging scarce labelled data to better support classification.

Our findings can help improve how children can access educational content online. In particular, we will explore the effectiveness of BiGBERT when applied to re-ranking search results on educational alignment as a step toward supporting search as learning among K-12 students [17, 31, 33]. BiGBERT provides a foundation to support research in other Information Retrieval areas, e.g., identification of resources that teachers may use in the classroom [10], automatic curation of resources for educational search engines similar to Infotopia [4], and identification of educational questions on question answering sites [34].

Acknowledgments. Work funded by NSF Award # 1763649. The authors would like to thank Dr. Ion Madrazo Azpiazu for his valuable feedback.

References

1. Abdessamed, O., Zakaria, E.: Web site classification based on url and content: Algerian vs. non-algerian case. In: Proceedings of the 12th International Symposium on Programming and Systems (ISPS). pp. 1–8. IEEE (2015)
2. Amazon, I.: Alexa top sites. https://www.alexa.com/topsites/category (2020), (accessed September 17, 2020)
3. Anuyah, O., Azpiazu, I.M., Pera, M.S.: Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain. In: Companion Proceedings of the World Wide Web Conference. pp. 274–282 (2019)
4. Bell, C., Bell, M.: Infotopia. https://wwww.infotopia.info (2020), (accessed August 17, 2020)
5. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. (2009)
6. Chen, W., Cai, F., Chen, H., De Rijke, M.: Personalized query suggestion diversification in information retrieval. Frontiers of Computer Science **14**(3), 143602 (2020)
7. Clavié, B., Gal, K.: Edubert: Pretrained deep language models for learning analytics. arXiv preprint arXiv:1912.00690 (2019)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Eickhoff, C., Serdyukov, P., de Vries, A.P.: Web page classification on child suitability. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 1425–1428 (2010)
10. Ekstrand, M.D., Wright, K.L., Pera, M.S.: Enhancing classroom instruction with online news. Aslib Journal of Information Management **72**(5), 725–744 (2020)
11. Elnaggar, A., Gebendorfer, C., Glaser, I., Matthes, F.: Multi-task deep learning for legal document translation, summarization and multi-label classification. In: Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference. pp. 9–15 (2018)
12. Francis, W.N., Kucera, H.: Brown corpus manual. Letters to the Editor **5**(2), 7 (1979)
13. Garbe, W.: Symspell. https://github.com/wolfgarbe/SymSpell (2020)
14. Geraci, F., Papini, T.: Approximating multi-class text classification via automatic generation of training examples. In: International Conference on Computational Linguistics and Intelligent Text Processing. pp. 585–601. Springer (2017)
15. Hashemi, M.: Web page classification: a survey of perspectives, gaps, and future directions. Multimedia Tools and Applications pp. 1–25 (2020)
16. Hassan, S., Mihalcea, R.: Learning to identify educational materials. ACM Transactions on Speech and Language Processing (TSLP) **8**(2), 1–18 (2008)
17. Hoppe, A., Holtz, P., Kammerer, Y., Yu, R., Dietze, S., Ewerth, R.: Current challenges for studying search as learning processes. Proceedings of Learning and Education with Web Data (2018)
18. Hughes, M., Li, I., Kotoulas, S., Suzumura, T.: Medical text classification using convolutional neural networks. Studies in Health Technology and Informatics **235**, 246–50 (2017)
19. Initiative, C.C.S.S.: Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects (2020), http://www.corestandards.org/wp-content/uploads/ELA_Standards1.pdf

20. Kastrati, Z., Imran, A.S., Yayilgan, S.Y.: The impact of deep learning on document classification using semantically rich representations. Information Processing & Management **56**(5), 1618–1632 (2019)
21. Liu, G., Guo, J.: Bidirectional lstm with attention mechanism and convolutional layer for text classification. Neurocomputing **337**, 325–338 (2019)
22. Nimmagadda, S.L., Zhu, D., Rudra, A.: Knowledge base smarter articulations for the open directory project in a sustainable digital ecosystem. In: Companion Proceedings of the International Conference on World Wide Web. pp. 1537–1545 (2017)
23. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the International Conference on Multimedia Information Retrieval. pp. 557–566 (2010)
24. Rajalakshmi, R., Aravindan, C.: A naive bayes approach for url classification with supervised feature selection and rejection framework. Computational Intelligence **34**(1), 363–396 (2018)
25. Rajalakshmi, R., Tiwari, H., Patel, J., Kumar, A., Karthik, R.: Design of kids-specific url classifier using recurrent convolutional neural network. Procedia Computer Science **167**, 2124–2131 (2020)
26. Rajalakshmi, R., Tiwari, H., Patel, J., Rameshkannan, R., Karthik, R.: Bidirectional gru-based attention model for kid-specific url classification. In: Deep Learning Techniques and Optimization Strategies in Big Data Analytics, pp. 78–90. IGI Global (2020)
27. Shen, D., Chen, Z., Yang, Q., Zeng, H.J., Zhang, B., Lu, Y., Ma, W.Y.: Web-page classification through summarization. In: Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 242–249 (2004)
28. Sreenivasulu, T., Jayakarthik, R., Shobarani, R.: Web content classification techniques based on fuzzy ontology. In: Intelligent Computing and Innovation on Data Science (Proceedings of ICTIDS 2019), pp. 189–197. Springer (2020)
29. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics. pp. 194–206. Springer (2019)
30. Tieleman, T., Hinton, G.: Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012)
31. Usta, A., Altingovde, I.S., Vidinli, I.B., Ozcan, R., Ulusoy, Ö.: How k-12 students search for learning? analysis of an educational search engine log. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1151–1154 (2014)
32. Xia, T.: Support vector machine based educational resources classification. International Journal of Information and Education Technology **6**(11), 880 (2016)
33. Yigit-Sert, S., Altingovde, I.S., Macdonald, C., Ounis, I., Ulusoy, Ö.: Explicit diversification of search results across multiple dimensions for educational search. Journal of the Association for Information Science and Technology (2020), online, https://doi.org/10.1002/asi.24403
34. Yilmaz, T., Ozcan, R., Altingovde, I.S., Ulusoy, Ö.: Improving educational web search for question-like queries through subject classification. Information Processing & Management **56**(1), 228–246 (2019)
35. Yu, S., Su, J., Luo, D.: Improving bert-based text classification with auxiliary sentence and domain knowledge. IEEE Access **7**, 176600–176612 (2019)

36. Zhao, W., Zhang, G., Yuan, G., Liu, J., Shan, H., Zhang, S.: The study on the text classification for financial news based on partial information. IEEE Access **8**, 100426–100437 (2020)