Towards Automatic, Scalable Quality Assurance in Open Education

Sahan Bulathwela[*] , **Emine Yilmaz** and **John Shawe-Taylor**
University College London
m.bulathwela@ucl.ac.uk

Abstract

With the emergence of Open Education Resources (OERs), educational content creation has boomed to a whole new scale. For AI-driven OER platforms such as X5GON, scalable quality assurance is highly impactful. As the quality of OERs could vary significantly, the quality assurance process plays a key role in maintaining a high-quality learner experience when using OERs. Managing this problem at large scale demands automating the quality assurance process as a whole or in parts. Prior research on automating quality assurance in the context of education is surprisingly scarce. We present our ongoing work of building Quality Assurance Models, a novel approach to using cross-modal features from OERs to predict quality using machine learning. While developing quality models, we extended our search beyond the education domain to identify features that indicate content quality that can be categorised into five main quality verticals. In the future, these features will enable us to leverage scalable quality assurance on OERs of different modalities. Furthermore, quality features will also become useful in learning quality preferences of learners when recommending content. Altogether, the expected outcomes of this research will mark a significant step towards Automatic, Scalable Quality Assurance in Open Education.

1 Introduction

Open Education Resources (OERs) can be defined as teaching, learning and research material that is available in the public domain or published under an open license. OERs can be of any medium and the open licensing allows anyone to consume, re-purpose and redistribute learning material with minimal costs and restrictions [UNESCO, 2019]. With the introduction of content creation strategies such as Content Explosion Model [Pawlowski *et al.*, 2007] and Open Educational Practice (OEP) [Ehlers *et al.*, 2018], new OERs are created on a day-to-day basis and the overall corpus grows rapidly. Through this work, we present our ongoing progress on building automatic, scalable quality models within the X5GON project (see section 1.1).

Progressing through this landscape, it is timely to find ways to facilitate quality assurance in the OER community as quality plays a critical role in the success of the movement. As the popularity of online learning has rapidly increased in recent years [Allen and Seaman, 2007], focusing on automatic, scalable quality assurance also creates an opportunity as most quality assurance solutions for educational material can be used across both OER and non-OER contexts. Although the reusability constraints vary, quality assurance also applies to commercially focused course creators who create **M**assive **O**pen **O**nline **C**ourses (MOOCs) and other educational materials.

1.1 X5GON Platform

OERs create social impact in developing and industrialised countries. X5GON (www.x5gon.org) is an international research collaboration, dedicated to the challenge of making OER more accessible, usable, reusable, and discoverable for educators and the general public. OERs come in various formats, including lecture videos and slides, e-books, tutorials, and podcasts. While the files remain with the provider, our platform extracts rich content representations which enable new possibilities for users to find information within the OER universe. X5GON envisages to develop and deploy highly impactful quality assurance models to improve adaptability of OERs by automatically identifying quality of materials at scale.

1.2 Outline

Firstly, in section 2, we explore how quality can be defined in education and how quality modeling is utilised in education. Secondly, in section 3, we identify what features indicate quality of content. Our search extends beyond the education domain because relevant work has been done on content quality in domains such as information retrieval. Thirdly, we outline the next steps of our ongoing work in section 4 and explain potential applications in section 5. Finally, we summarise our progress in section 6.

[*] Contact Author

2 What is Quality in Open Education?

[Camilleri *et al.*, 2014] identifies several quality related issues in the context of Open Educational Resources. In the above mentioned report, the need for investing more research into creating standards and quality enforcement tools for education resources is highlighted repeatedly.

Before we dive into addressing scalable quality assurance in education, it is imperative that we understand what quality means in the context of open education. A high quality educational material is an educational resource that enables the the learner to achieve the expected learning outcomes.

[Lane, 2010] argues that designing effective educational content entails satisfying three main features:

1. Material being academically sound in that it appropriately covers the body of knowledge and meaning for the topic.
2. Material being pedagogically robust in that the way the material is structured matches a stated pedagogical model and sets out appropriate learning outcomes and ways of assessing those outcomes.
3. Material is presented through appropriate choices of media that are helpful for learners to meet the learning outcomes.

Lane further argues that *Learner-Content interaction* heavily influences the time and effort an individual learner will commit to achieving the given or self-set learning goals. The above arguments suggest that a high quality educational material comprises features such as facilitating accuracy, pedagogical robustness, and engageability.

Quality assurance approaches can be categorised based on their scope. [Clements and Pawlowski, 2012] outlines three categories of quality approaches in education.

1. ***Generic Quality Approaches*** that refer to concepts and procedures providing quality management in general, independent of the domain.
2. ***Specific Quality Approaches*** that refer to standards and mechanisms providing quality management in the domain of Technology Enhanced Learning.
3. ***Specific Quality Instruments*** that refer to standards, tools and mechanisms providing quality management related to specific purposes and functions.

2.1 Challenges in Quality Assurance of OERs

As observed from the previous section, quality of educational material is a multi-faceted problem. The quality of an educational resource depends on different aspects, such as the accuracy of content, structuring of materials and the engageability of a resource.

Although these features can be successfully managed using standard quality assurance mechanisms such as ISO standards [ISO, 2017], these approaches are often hard to scale up. Enforcing standards may seem feasible at organisational level. However, it is quite difficult to force informal resource creators (the life source of the OER movement) to adhere to such standards. In the context of OERs, it is more difficult due to the wider flexibility for the authors to reuse and re-purpose educational material [Clements and Pawlowski, 2012].

A key challenge in applying AI for scalable quality assurance in education is the scarcity of available and labelled datasets. Labelled quality datasets of educational materials are very hard to come by although there are a handful of datasets that are aimed to support similar tasks. As observed in other fields [Boyer *et al.*, 2017; Pitler and Nenkova, 2008], getting experts to annotate the quality of educational resources is an option. However, labelling educational material is time consuming and carries huge opportunity costs.

2.2 Machine Learning Models and Quality Assurance in Education

In the education sector, statistical modelling has made prominent contributions in the area of educational economics [Marschark *et al.*, 2015; Weerahewa *et al.*, 2013]. When we consider how AI and ML can assist in the quality assessment of OERs, we face the reality that that some of the quality aspects outlined in section 2 are far from automation at this point. For instance, automatically measuring the correctness or academic soundness of material is quite ambitious. Although aspects such as argument strength measurement [Persing and Ng, 2015] could help, it doesn't guarantee correctness. Fact checking and misinformation detection is an actively researched area where datasets and challenges have only started appearing recently [Thorne *et al.*, 2018].

In the context of predicting quality, Automatic Essay Scoring (AES) [Islam and Hoque, 2010; Yannakoudakis *et al.*, 2011] and Quality Assessment of Online Digital Libraries such as Wikipedia [Dalip *et al.*, 2011] are two relevant strands of research exploring how AI can be used for quality assessment in education. It is evident that the majority of features used for quality prediction in the above cases focus on presentation, structure and other engagement related aspects of quality.

In essence, this shows that the ideal approach at this point is to focus on *Specific Quality Instruments* (See section 2) that focus on engagement aspects of educational resources.

3 Factors Impacting Quality of Content

In the information era, assessing the quality of content is not just a concern of the education domain. We learn that multiple domains beyond education carry out active research into automatically assessing quality of information. For instance, modelling trustability of web forums is an area of active research in the healthcare domain that emphasises on quality of the posts. Document quality assessment is quite important in information retrieval domain as well. By doing an extensive literature survey, we identified several factors that indicate quality of a document. Five main *quality verticals* seem to emerge consistently across multiple research domains when quality is discussed. Namely, *understandability*, *topic coverage*, *freshness of information*, *presentation*, and *authority*. Figure 1 summarises the features we identified along with the quality verticals they belong to.

In the following we detail the different factors we observed to affect the quality of information.

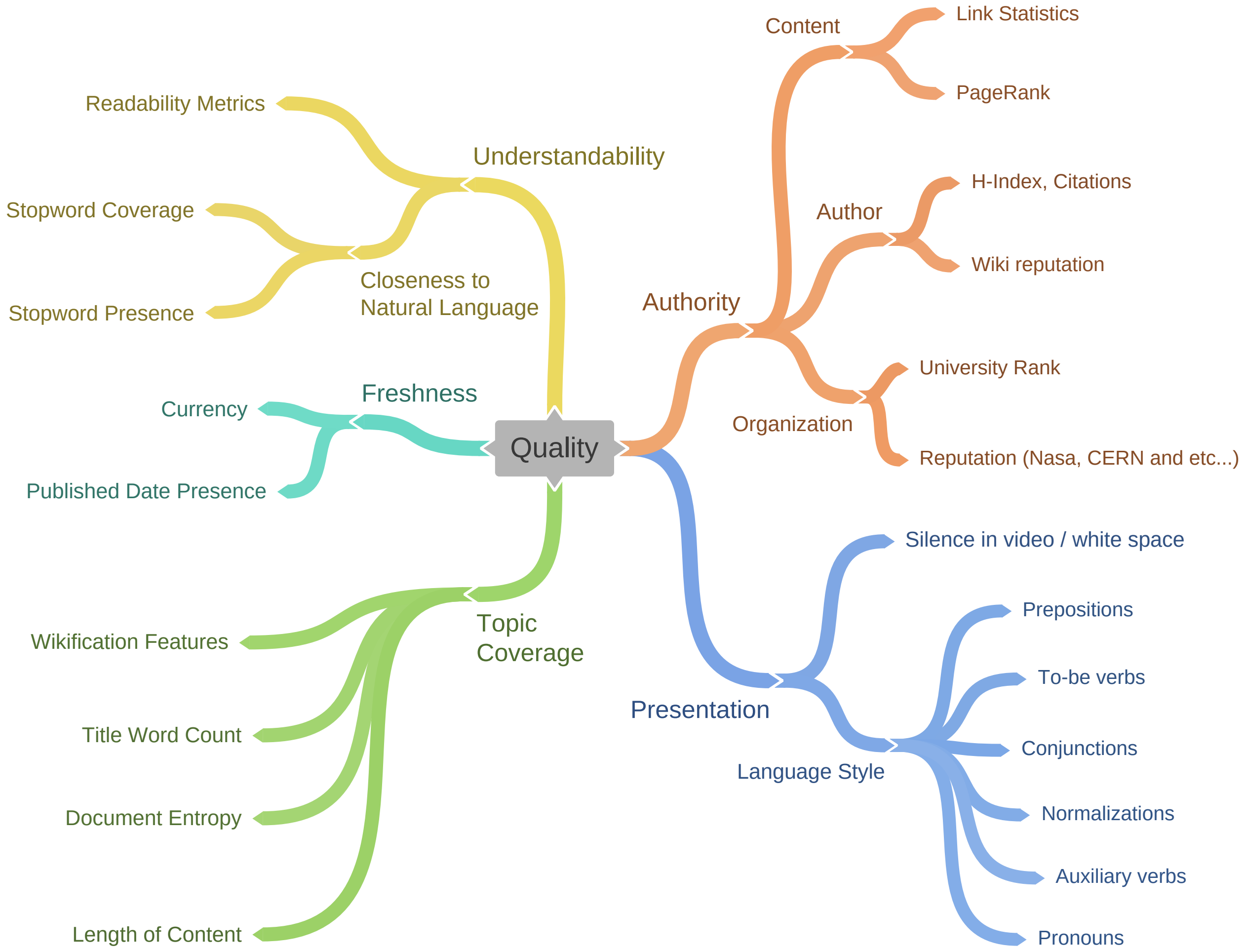

Figure 1: Five Quality Verticals: *Understandability (Yellow)*, *Topic Coverage (Green)*, *Freshness of Information (Cyan)*, *Presentation (Blue)*, and *authority (Orange)*

3.1 Understandability

Understandability of content mainly relates to the *effort* of learners in consuming the material. Metrics such as Fletch Kuncaid Score (FKS), Fletch Readability Ease Score (FRES), Gunning fog index (FOG), and Simple Measure of Gobbledygook (SMOG) have emerged from the scientific community and they are widely adopted when measuring the readability level of documents [Si and Callan, 2001]. In educational search, incorporating the readability level of documents has shown to improve the relevance of documents retrieved for a learner [Collins-Thompson *et al.*, 2011]. By accounting for student effort, [Syed and Collins-Thompson, 2017] further showed that information search for learning can be improved.

While level of language affects the understandability of material, the style of language used heavily impacts the information delivery. We found numerous studies that use features such as the intersection between English stop-words and document vocabulary [Ntoulas *et al.*, 2006], the deviation of word distribution from a typical document [Zhou and Croft, 2005] to represent the style of language used in content.

3.2 Topic Coverage

[Syed and Collins-Thompson, 2017] have successfully improved information search results for learning tasks by using the coverage of knowledge components and topics to represent documents. Their method primarily represents the document as a distribution of knowledge components. However, identifying knowledge components or topics in documents poses a challenge as unsupervised topic detection techniques such as LDA [Blei *et al.*, 2003] don't guarantee optimal results. Wikification, a more recent approach, can extract Wikipedia topics related to a text with meta information about how these topics are linked together in Wikipedia [Brank *et al.*, 2017]).

In the information search domain, document entropy has been used to quantify the focus of a document. This measure determines whether a document is narrowly focused on a few topics or widely discusses a range of topics. [Bendersky *et al.*, 2011] uses document entropy as a feature in modelling quality biased information search.

Length features of text documents, such as document length and title length, have consistently proved to be useful in predicting quality [Ntoulas *et al.*, 2006; Dalip *et al.*, 2011].

3.3 Freshness of Information

Validity of information may decay over time. In the context of healthcare forums, having the publication date mentioned is considered a good feature of quality content. *date extracts* have been used to automate publication date detection in health forums [Boyer *et al.*, 2017] using Named Entity Recognition (NER).

Zhu and Gauch refer to *currency* that aims to measure how current the information on a website is. They saw that the search effectiveness improved when they incorporated currency in search engines [Zhu and Gauch, 2000]. Information age is also used in evaluating quality of Wikipedia articles [Warncke-Wang *et al.*, 2013].

3.4 Presentation Aspects

The nature of presenting information is also an important feature when deciding on the quality of an information resource. Features such as percentage of coherent text indicates if text is scattered around the web page with advertising spaces in between. [Sondhi *et al.*, 2012] captures how white space is spread across a web page to represent how well information is presented [Warncke-Wang *et al.*, 2013]. When talks or lectures are considered, the flow of the talk heavily depends on pauses and where breaks are used.

Furthermore, word groups such as pronouns, conjunctions etc. heavily impact the language style and show to be helpful in automatic quality assessment [Dalip *et al.*, 2011].

3.5 Authority of Content

In education, authority is taken very seriously. When asked, 55% of teachers who create courses were found to believe that high quality material comes from reputable sources such as CERN, Harvard and NASA [Clements and Pawlowski, 2012].

Quality frameworks such as HONCode treat authority as a core component. The qualifications of the health forum authors and their affiliations give a huge weight towards the reliability of information [Boyer and Dolamic, 2014]. Link structures (an indicator of content authority) have been used for trustability evaluation as well [Sondhi *et al.*, 2012].

4 Next Steps

The sensible immediate next step for us is to use the identified quality features in training quality models using machine learning. A suitable dataset has already been extracted from a popular OER repository, VideoLectures.Net (VLN)[1]. Once a model has been trained, we would identify what quality features are most impactful when predicting quality of an OER.

4.1 Ongoing Challenges

The VLN dataset we have identified comes with certain caveats. It is a collection of video lectures. Most of the lectures are in English. These attributes may expose the trained models to modality and language biases. We would want to validate the generalisability of the quality models across different modalities and languages.

Another challenge we face is how to capture more generalisable features that represent authority of content. Most author authority features that we identified in section 3 are highly domain specific (e.g. citation indices of scientific authors and reputation of organisations such as MIT, CERN etc). In the context of OERs, authors may not necessarily inherit such domain specific credentials which leads to finding more general, but useful authority features.

Another challenge is to understand how to communicate the quality predictions to the relevant stakeholders (learner, teacher, creator and etc...) in a highly interpretable manner. This will improve transparency of the document quality and also improve user perception of quality. What approach to take in interfacing the quality predictions with the learner in a more descriptive manner is still an open question.

[1] www.videolectures.net

5 Beneficial Applications

We see two main applications in the education domain that would strongly benefit from the identified quality features. Namely, (1) Automatic, Scalable Quality Assurance, (2) Personalised Recommendation.

Automatic, Scalable Quality Assurance The quality features we identify could be used to derive a quality score that can enrich learning materials with quality related information. This quality score can be derived by building quality models that use the above features. When integrated into future user interfaces, this information can potentially help learners and teachers make more informed choices of educational material. While existing quality assurance frameworks [ISO, 2017] can provide written guidelines for content creators, our research enables new kinds of use cases. Particularly, it could be purposed into a realtime support tool, providing instant feedback to a content creator. By providing transparency about which features lead to a higher quality score, the system could help authors continuously improve the quality of content during production and reuse. At scale, this could encourage continuous improvement of quality in OER. Administrators that oversee massive repositories such as X5GON are also empowered with quality instruments that can evaluate quality of educational materials at scale.

Personalised Recommendation Furthermore, recommendation algorithms such as [Covington *et al.*, 2016] can benefit from using the quality features identified in our analysis. By using them to learn specific quality preferences of individual learners, personalisation algorithms can learn what quality choices learners make based on how they engage with different learning resources.

6 Summary

We present the ongoing work of building quality assurance models for OERs. By surveying different research domains, we identify numerous quality features that can be categorised into five main quality verticals. We have also identified a suitable OER dataset (VLN dataset) that will enable extracting the above features and building quality models. We have identified several challenges relating to the potential quality models that can be developed using the VLN dataset. Once these models are built and the challenges are overcome and validated, outcomes of this work can be used in evaluating one aspect of quality of educational resources, namely, *engagement quality*. The identified quality features can also be used as part of personalisation algorithms by encoding user quality preferences using the above features. Altogether, these outcomes will mark a significant step towards Automatic, Scalable Quality Assurance in Open Education.

Acknowledgements

The research described in this paper was conducted as part of the X5GON project (www.x5gon.org) which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 761758.

References

[Allen and Seaman, 2007] I. Elaine Allen and Jeff Seaman. Online nation: Five years of growth in online learning. Technical report, oct 2007.

[Bendersky *et al.*, 2011] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. Quality-biased ranking of web documents. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 95–104, New York, NY, USA, 2011. ACM.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[Boyer and Dolamic, 2014] Célia Boyer and Ljiljana Dolamic. Feasibility of automated detection of honcode conformity for health-related websites. *International Journal of Advanced Computer Science and Applications*, 5(3), 2014.

[Boyer *et al.*, 2017] Célia Boyer, Cédric Frossard, Arnaud Gaudinat, Allan Hanbury, and Gilles Falquetd. How to sort trustworthy health online information? improvements of the automated detection of honcode criteria. *Procedia Computer Science*, 121:940 – 949, 2017.

[Brank *et al.*, 2017] Janez Brank, Gregor Leban, and Marko Grobelnik. Proceedings of the slovenian conference on data mining and data warehouses. SiKDD '17, 2017.

[Camilleri *et al.*, 2014] Anthony F. Camilleri, Ulf Daniel Ehlers, and Jan Pawlowski. *State of the art review of quality issues related to open educational resources (OER)*, volume 52 S. - JRC Scientific and Policy Reports of *Publications Office of the European Union 2014*. Luxembourg, 05/2014 2014.

[Clements and Pawlowski, 2012] K.I. Clements and J.M. Pawlowski. User-oriented quality for oer: understanding teachers' views on re-use, quality, and trust. *Journal of Computer Assisted Learning*, 28(1):4–14, 2012.

[Collins-Thompson *et al.*, 2011] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 403–412, New York, NY, USA, 2011. ACM.

[Covington *et al.*, 2016] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 191–198, New York, NY, USA, 2016. ACM.

[Dalip *et al.*, 2011] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Automatic assessment of document quality in web collaborative digital libraries. *J. Data and Information Quality*, 2(3):14:1–14:30, December 2011.

[Ehlers *et al.*, 2018] Max Ehlers, Robert Schuwer, and Ben Janssen. Oer in tvet: Open educational resources for skills development, 2018.

[Islam and Hoque, 2010] M. Monjurul Islam and A. S. M. Latiful Hoque. Automated essay scoring using generalized latent semantic analysis. In *2010 13th International Conference on Computer and Information Technology (ICCIT)*, 2010.

[ISO, 2017] ISO. *ISO/IEC 40180:2017(en): Information technology — Quality for learning, education and training — Fundamentals and reference framework*, 2017. Accessed: 2018-11-03.

[Lane, 2010] Andy Lane. Open information, open content, open source. In Richard N. Katz, editor, *The Tower and The Cloud*, pages 158–168. EDUCAUSE, USA, 2010.

[Marschark *et al.*, 2015] Marc Marschark, Debra M. Shaver, Katherine M. Nagle, and Lynn A. Newman. Predicting the academic achievement of deaf and hard-of-hearing students from individual, household, communication, and educational factors. *Exceptional Children*, 81(3):350–369, 2015.

[Ntoulas *et al.*, 2006] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 83–92, New York, NY, USA, 2006. ACM.

[Pawlowski *et al.*, 2007] Jan M. Pawlowski, Volker Zimmermann, and Imc Ag. Open content: A concept for the future of e- learning and knowledge management?, 2007.

[Persing and Ng, 2015] Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 543–552, 2015.

[Pitler and Nenkova, 2008] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 186–195, 2008.

[Si and Callan, 2001] Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA, 2001. ACM.

[Sondhi *et al.*, 2012] Parikshit Sondhi, V. G. Vinod Vydiswaran, and ChengXiang Zhai. Reliability prediction of webpages in the medical domain. In Ricardo Baeza-Yates, Arjen P. de Vries, Hugo Zaragoza, B. Barla Cambazoglu, Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri, editors, *Advances in Information Retrieval*, pages 219–231, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[Syed and Collins-Thompson, 2017] Rohail Syed and Kevyn Collins-Thompson. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 555–564, New York, NY, USA, 2017. ACM.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

[UNESCO, 2019] UNESCO. Open educational resources (oer), 2019. Accessed: 2019-04-01.

[Warncke-Wang *et al.*, 2013] 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.

[Weerahewa *et al.*, 2013] Jeevika Weerahewa, Sahan Bulathwela, Pradeep Silva, and Kalyani Perera. An analysis of academic performance of undergraduates: Effects of academic vis-a-vis non-academic factors. 2013.

[Yannakoudakis *et al.*, 2011] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Zhou and Croft, 2005] Yun Zhou and W. Bruce Croft. Document quality models for web ad hoc retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 331–332, New York, NY, USA, 2005. ACM.

[Zhu and Gauch, 2000] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 288–295, New York, NY, USA, 2000. ACM.