MACHINE LEARNING FOR QUALITY EVALUATION OF WIKIPEDIA STUDENT WRITINGS

Hector Gabriel Corrale de Matos
University of São Paulo

Alexandre Alberto Pascotto Montilha
University of São Paulo

Thais Catalani Morata
National Institute for Occupational Safety and Health

Ana Paula Berberian
University Tuiuti do Paraná

Lilian Cássia Bornia Jacob
University of São Paulo

Abstract

The aim of this abstract is to assess the Wikipedia article quality prediction model based on the Lift Wing-ORES machine learning infrastructure as a methodology for evaluating student writing on Wikipedia.

Keywords: Machine Learning, Wikipedia, Education, Writing Assessment, Lift Wing

Introduction

Quality assessment in peer production projects, such as Wikipedia, is crucial for understanding project dynamics and ensuring content reliability. Traditional methods rely on human evaluation, which is time-consuming and subject to limited perspectives (Kane and Ransbotham, 2016). Machine learning (ML) have been proposed to address these challenges. The Objective Revision Evaluation Service (ORES)1 uses a tree-based classifier to predict the quality class of Wikipedia articles (Halfaker and Geiger 2019). This model was trained on existing quality assessments and estimated probability outputs for each quality class. The model employs a sklearn library to fit classifiers by minimizing multinomial deviance. For each Wikipedia article with predictors and labeled quality class, the model estimates the probabilities for each quality class (Stub, Start, C-class, B-class, Good Article, Featured Article). These probabilities sum to one, yielding a unit vector for each article. The model calculates the loss based on the difference between the predicted and true quality class probabilities (TeBlunthuis, 2018). Using supervised ML algorithms, ORES predicts the quality of an edit based on various characteristics, such as the amount of text added or removed, presence of links, and grammatical quality, among others. The ORES model demonstrates the potential of ML to extend quality measurements in peer production projects. Its application highlights both the benefits and challenges associated with automated quality assessment in collaborative environments such as Wikipedia. ORES is being replaced by Lift Wing, a more versatile ML platform to extract quantitative data from Wikimedia content2. Wikipedia is used as an active teaching methodological tool in undergraduate and graduate coursework during editing campaigns and edit-a-thons. Students were assigned to edit and enhance the content of Wikipedia articles in terms of information accuracy, reliance on reliable sources, depth of knowledge, and grammatical and lexical quality of the texts. Working with the digitally disseminated textual genre and collaborative writing medium of Wikipedia articles provides instructors and researchers the opportunity to measure the enhancement in textual quality resulting from student edits (Montilha et al. 2023). The assessment of students’ contributions to Wikipedia is typically conducted by teachers or supervisors through structured rubrics and evaluation criteria. However, this model presents relative subjectivity and requires considerable time for evaluation. The use of objective quantitative measures to analyze the textual quality of Wikipedia articles by employing ML models as an educational assessment strategy for student edits on Wikipedia has the potential to improve and expedite the process. This study aimed to explore the effectiveness of ORES in comparison to human-based evaluation for the assessment of the quality of Wikipedia articles.

Methods

Undergraduate students from a Brazilian university were assigned to edit Wikipedia for coursework on hearing health in 2023. They received training in article editing, covering editing norms, the use of bibliographic references, and multimedia resources to enhance content related to hearing health. The students’ edits and articles on Portuguese Wikipedia were tracked using the Programs & Events Dashboard platform3. The course resulted in 35 Portuguese

1https://w.wiki/49jW
2https://w.wiki/9Uy6
3https://outreachdashboard.wmflabs.org/courses/USP/Teoria_e_Diagnóstico_Audiológico_II
articles, 5 new ones created and 30 edited by the students. The methodology was implemented in two stages: (i) Machine Evaluation, and (ii) Human Evaluation. Both stages were carried out prior to and following the students’ edits. This approach allowed for a comprehensive assessment of the students’ contributions.

**Machine Evaluation**: The ORES ptwiki-articlequality model was applied to 23 edited articles based on revisions before and after the students’ edits. The model does not evaluate the quality of writing, but the structural characteristics of articles that correlate with good writing. The model outputs scores for article quality on Portuguese Wikipedia ranging from 1 (draft text) to 6 (complete text). This stage was executed using Python via a Jupyter notebook on Wikimedia Web Shell (https://public-paws.wmcloud.org/User:CorraleH/WikiWorkshop24articlequality.ipynb).

**Human Evaluation**: Three articles were randomly selected for evaluation by course instructors using a structured evaluation rubric. The rubric used in the assessment was translated into Portuguese from the material "An example grading rubric for evaluating students’ Wikipedia article writing assignments" (Figure 1) produced by the Wiki Education Foundation. The rubric results were quantitative, with a total score of 45 points, aligned with Wikipedia’s style guide and writing rules. This stage was conducted using an electronic form and tabulated for descriptive analysis. Only existing articles were analyzed using both evaluation models.

**Results**

The Machine Evaluation (Table 1) revealed an average quality improvement of 33.3%, reflecting a general gain of three to four points. Substantial enhancements were observed, such as a 400% increase (from 1 to 5 points) in the article titled "Neuroma do Acústico" (Acoustic Neuroma) and a 200% rise (from 1 to 3 points) in the article addressing the “Aparelho Vestibular” (Vestibular System). These findings underscore the effectiveness of student-led editing efforts in improving content quality. Across the board, all articles exhibited either an improvement or, at the very least, a maintenance of quality after the editing by the students. The outcomes of the Human Evaluation (Table 2), conducted utilizing a structured rubric, echoed the trend of quality enhancement observed in the articles, as demonstrated by the Machine Evaluation model.

**Discussion/Conclusions**

Wikimedia platforms have been used in various educational contexts such as university courses and community projects. These projects, which are characterized as collaborative knowledge-building tools, contribute to open science models. Wikipedia facilitates the observation of practices and concepts related to writing processes, scientific research, collaboration, and narrative construction, thus enabling the authentic dissemination of students' textual production experiences. Leveraging ML models can benefit the assessment of textual quality, making the evaluation of student activities on Wikipedia more efficient (Bernius, Krusche, and Bruegge 2022). Using Wikipedia in academic instruction can enhance digital literacy and active learning methodologies, improving students’ scientific writing skills. Written content produced for wide-reaching audiences promotes reading and writing practices and advances the use of text in academic contexts. Evaluating Wikipedia as a literacy tool and educational resource is necessary to develop writing skills in undergraduate students and other educational levels. Automating quality assessment with ML models provides objective measures for improvement in Wikipedia articles edited by students, allowing more participants and increasing the quality of information in specific fields. However, human evaluations by teachers offer deeper assessments of knowledge quality, reference sources, and narrative construction.

Further research is required to develop joint methodologies that combine machine-based and human-centered approaches for scalable assessment structures. ML in educational assessments must be ethically guided by human oversight to ensure quality and integrity. Future studies should address the subjectivity in human assessments, develop robust systems, and provide comprehensive understanding of ORES’s implications in Portuguese comparable to English Wikipedia.

**References**


---

4https://w.wiki/9cMH
Table 1: ORES results before (B) and after (A) the edits made by students on Portuguese Wikipedia. Higher numbers indicate higher-quality scores. The article quality on Portuguese Wikipedia ranges from 1 (draft text) to 6 (complete text).

<table>
<thead>
<tr>
<th>Article</th>
<th>B</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geriatria</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Psicoacústica</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mascaramento em audiologia</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Perda auditiva induzida por ruído</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Doença de Ménière</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Audiometria</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Deficiência auditiva</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Aparelho vestibular</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Audiograma</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Audição</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Neura do acústico</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Habilitação e reabilitação auditiva</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: Human evaluation based on rubric compared with ORES results before (B) and after (A) the edits made by students on Portuguese Wikipedia. * The maximum rubric score was 45. Estimated value by the model ptwiki-articlequality: predicted and probability value of estimation accuracy (0 to 1).

![Wiki Workshop](https://w.wiki/8Xzv) produced by the Wiki Education Foundation, translated into Portuguese (https://w.wiki/8$U6)