

# 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)<sup>1</sup> 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 content<sup>2</sup>. 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* platform<sup>3</sup>. The course resulted in 35 Portuguese

<sup>1</sup><https://w.wiki/49jW>

<sup>2</sup><https://w.wiki/9Uy6>

<sup>3</sup>[https://outreachdashboard.wmflabs.org/courses/USP/Teoria\\_e\\_Diagnóstico\\_Audiológico\\_II](https://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*<sup>4</sup> 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

- J. P. Bernius, S. Krusche, and B. Bruegge, "Machine learning based feedback on textual student answers in large courses," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100081, 2022, doi: 10.1016/j.caeai.2022.100081.
- A. A. P. Montilha, T. C. Morata, D. Á. Flor, M. A. A. M. Machado, F. A. Menegon, and F. Zucki, "The Promotion of Hearing Health through Wikipedia Campaigns: Article Quality and Reach Assessment," *Healthcare*, vol. 11, no. 11, p. 1572, May 2023, doi: 10.3390/healthcare11111572.
- G. C. Kane and S. Ransbotham, "Research Note—Content and Collaboration: An Affiliation Network Approach to Information Quality in Online Peer Production Communities," *Information Systems Research*, vol. 27, no. 2, pp. 424–439, Jun. 2016, doi: 10.1287/isre.2016.0622.
- A. Halfaker and R. S. Geiger, "ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia." arXiv, 2019. doi: 10.48550/ARXIV.1909.05189.
- N. Teblunhuis, "Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression," in *17th International Symposium on Open Collaboration*, Online Spain: ACM, Sep. 2021, pp. 1–10. doi: 10.1145/3479986.3479991.

<sup>4</sup><https://w.wiki/9cMH>

Article	B	A	Article	B	A
Geriatrics	2	4	Paralisia facial	2	3
Psicoacústica	3	3	Otosclerose	2	3
Mascaramento em audiologia	1	2	Perda auditiva condutiva	2	3
Perda auditiva induzida por ruído	4	6	Efeitos da poluição sonora na saúde	3	3
Doença de Ménière	5	5	Presbiacusia	5	5
Audiometria	2	3	Ototoxicidade	4	4
Deficiência auditiva	4	4	Reflexo Acústico	5	5
Aparelho vestibular	1	3	Dia Mundial da Audição	3	5
Audiograma	2	3	Acufeno	4	4
Audição	4	4	Dia Internacional da Conscientização Sobre o Ruído	3	3
Neuroma do acústico	1	5	Percepção auditiva	4	4
Habilitação e reabilitação auditiva	3	3	-	-	-

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).

Articles selected on Wikipedia for human evaluation	B - Quality prior to editing by students		A - Quality after editing by students	
	Rubric	ORES*	Rubric	ORES*
Otosclerose	22	2 (0.339)	35	3 (0.285)
Neuroma do acústico	10	1 (0.653)	43	5 (0.253)
Doença de Ménière	28	5 (0.268)	39	5 (0.276)

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).

Wikipedia Assignment Assessment

A guide for evaluating student contributions to Wikipedia.

	Excellent	Good	Fair	Poor	
1. Lead Section	<b>Introductory sentence</b> 1.1	States article topic concisely and accurately in single sentence	Topic of article stated, though not concise/direct.	Begins with an introduction, not a lead	No lead
	<b>Summary</b> 1.2	Summarizes all major points in the article	Summarizes most major points, but misses one or more important aspects	Includes excessive background information	Summary missing, lacking key ideas
	<b>Context</b> 1.3	All information included is also present in body of the article	Includes some information not present in body of the article	Includes only 1-2 additional sentences of information	Doesn't provide enough information to determine what the article is about
Points: <input type="text"/>					
2. Article	<b>Organization</b> 2.1	Clear organization of heading and subheadings, appropriate transitions and clear language/grammar	Purposeful organization, but article does not flow between sections	Confusing organization and/or many grammatical errors	No sections
	<b>Content</b> 2.2	Covers info relevant to assigned topic; links to relevant articles for background	Covers most of the assigned topic area	Covers some of the assigned topic area	Misses the point
	<b>Balance</b> 2.3	Article presents balanced coverage without favoring one side unduly	Article presents one side, ignores minority views	Article attempts to convince readers of majority view	Article presents fringe view as if it were mainstream
	<b>Tone</b> 2.4	Tone is neutral and appropriate for an encyclopedia audience	Tone is mostly good, but becomes informal or chatty in places	Content appeals to the reader directly (uses you, I, we, or one)	Additions are promotional
	<b>Images</b> 2.5	Images improve the reader's understanding of the topic. Captions are clear, concise.	Images are relevant. Article is more visually attractive. Captions are too detailed.	No images, or images of limited relevance. Captions are absent or confusing.	Irrelevant images. Images that break the layout of the page. Copyright violations.
Points: <input type="text"/>					
3. References	<b>Citations</b> 3.1	Every statement can easily be associated with a supporting reference	A few statements at the end of some paragraphs have unclear sourcing	A few unsourced paragraphs or sections	Very few or no sources
	<b>Sources</b> 3.2	Most sources are the best available, are appropriate for the discipline/genre	Article uses mostly good sources, but includes some lower-quality sources	Article depends heavily on non-independent sources or uses many low-quality sources	Article uses unreliable internet sources
	<b>Completeness</b> 3.3	Most references include completely filled-out citation template or are otherwise complete	Most references are fairly complete, but some are missing something	References have enough information to track down sources, but with difficulty	References lack important information; sources are too hard to track down
Points: <input type="text"/>					

Figure 1: English version of "An Example Grading Rubric for Evaluating Students' Wikipedia Article Writing Assignments," (<https://w.wiki/8Xzv>) produced by the Wiki Education Foundation, translated into Portuguese ([https://w.wiki/8\\$U6](https://w.wiki/8$U6))