Automatic Quality Assessment of Wikipedia Articles - A Systematic Literature Review

Journal:	*Computing Surveys*
Manuscript ID	CSUR-2022-0228.R1
Paper:	Long Survey Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Moás, Pedro; Universidade do Porto Faculdade de Engenharia, Lopes, Carla; Universidade do Porto Faculdade de Engenharia; INESC TEC
Computing Classification Systems:	Information systems → Wikis, Computing methodologies → Natural language processing, Computing methodologies → Machine learning, Applied computing → Document management and text processing

Automatic Quality Assessment of Wikipedia Articles - A Systematic Literature Review

PEDRO MIGUEL MOÁS, Faculdade de Engenharia da Universidade do Porto, Portugal
CARLA TEIXEIRA LOPES, Faculdade de Engenharia da Universidade do Porto, INESC TEC, Portugal

Wikipedia is the largest online encyclopedia in the world. Its collaborative nature makes it challenging to ensure article quality. Wikipedia designed a quality scale, but that assessment process is primarily manual, so many articles remain unassessed. We review existing methods for automatically measuring the quality of Wikipedia articles, identifying and comparing machine learning algorithms, article features, quality metrics, and used datasets. We included 149 distinct studies, hundreds of features, and many results from numerous machine learning experiments. This review serves as a basis for future work on the automatic assessment of Wikipedia, exploring common themes and gaps in the literature.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Information systems** → **Wikis**; • **Applied computing** → *Document management and text processing*.

Additional Key Words and Phrases: Wikipedia, Quality Assessment, Information Quality

ACM Reference Format:
Pedro Miguel Moás and Carla Teixeira Lopes. 2023. Automatic Quality Assessment of Wikipedia Articles - A Systematic Literature Review. *J. ACM* 37, 4, Article 111 (August 2023), 37 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Wikipedia is the largest and most well-known online encyclopedia and has kept its growing pace for years. As of April 2023, it contains over 6.6 million English articles [164], with versions across a list of 321 active languages [161].

Not only is Wikipedia free, but it is also fully managed by human volunteers, averaging contributions at a rate of 5.7 edits per second during 2022 [160]. Its reader base is growing steadily, considering that, last year, Wikipedia totaled close to 280 billion page views across 2 billion unique devices [160]. Some studies even show that most search engines frequently include Wikipedia pages in their results. According to Vincent and Hecht [154], 80% of *common* (frequent) queries and 70% of *trending* queries (news/events) return results from Wikipedia on the first page, when tested with search engines like Google, Bing, and DuckDuckGo.

The fully collaborative aspect of Wikipedia brings its own set of challenges too, as the lack of centralized authority over the editors makes it challenging to ensure quality throughout the website. Only 8.1% of edits are reverted [160], showing that vandalism and the so-called *revert wars* are relatively uncommon, but it is still crucial to ensure the improvement of low-quality articles.

Authors' addresses: Pedro Miguel Moás, up201705208@edu.fe.up.pt, Faculdade de Engenharia da Universidade do Porto, R. Dr. Roberto Frias, s/n, Porto, Portugal, 4200-465; Carla Teixeira Lopes, ctl@fe.up.pt, Faculdade de Engenharia da Universidade do Porto, INESC TEC, R. Dr. Roberto Frias, s/n, Porto, Portugal, 4200-465.

Another issue is the substantial quality discrepancy between English Wikipedia and its other versions. First, each non-English version covers a much smaller amount of articles [161]. Also, they are often much more incomplete as well. Roy et al. [115, 116] demonstrate that English articles from Wikipedia are usually longer than their translations. They determined that German articles are, on average, 30% shorter than English ones, and for Spanish articles, that value increases to 47%, but there are still some English articles that are much less complete than their non-English counterparts. Couto and Lopes [17] have also shown this quality discrepancy, although only focused on health-related articles. They used a set of metrics to determine that English articles show the best values for quality, ranking much higher than other idioms.

To better monitor and help maintain the quality of the website, Wikipedia designed a quality scale that aims to rate articles within one of 9 possible grades, which go from the most incomplete documents (Starts and Stubs) to the most comprehensive, well-written articles (Featured Articles) [163]. However, the majority of Wikipedia is made up of lower-quality articles, with Starts and Stubs accounting for more than 80% of its English content. In comparison, the share of Featured Articles and Good Articles is around 0.7% [163]. Nonetheless, these values are not meant to be taken as official ratings but instead for internal use by the contributors. Besides, not every English article is rated, and the non-English versions of Wikipedia that also assess their content will have different quality scales, evaluated with other criteria. For those reasons, Wikipedia users lack a consistent and transparent method for determining the quality of articles.

Our goal is to review proposed methods for automatically measuring the quality of Wikipedia articles. Hence, we conduct a systematic literature review to assess the state-of-the-art within this topic, examining and comparing existing approaches used to automatically measure article quality. Specifically, we analyzed machine learning methods, article features, quality metrics, datasets, and other common aspects of these approaches, such as multilingual assessment and data visualization/explanation tools for supporting the reader and editor community. With this review, we aim to provide a starting point for future work that aims to understand Wikipedia quality and design automatic methods for measuring it.

We divided this article into eight sections. After this introduction in Section 1, Section 2 provides some insight about Information Quality. Section 3 details our methodology for the systematic review, and we list its results in the following sections: Section 4 overviews the included papers and the methods they use, Section 5 summarizes used machine learning approaches, and Section 6 analyses applied article features and quality metrics. We summarize and discuss our findings in Section 7, answering the defined research questions. Finally, we conclude this paper in Section 8, where we reflect on our study and examine future work possibilities.

2 INFORMATION QUALITY

It is important to design our definition of quality. Hence, we must first answer: *What is quality? Can it be objectively quantified?* Information Quality (IQ) is an extraordinarily researched topic, and there exist numerous attempts to provide a way to calculate it. Lee et al. [80] break down the measurement of Information Quality into 15 properties, including *Accessibility*, *Believability*, *Interpretability*, *Objectivity*, *Reputation*, and *Timeliness*. Some of these aspects are much easier to assess than others. However, with recent developments in Natural Language Processing (NLP), some works already attempt to evaluate more complex topics such as bias [65], neutrality [73], trustworthiness [2, 38, 74, 76, 182]. Although these studies are an inspiration for authors attempting to tackle the topic of this review, we do not intend to include them unless they specifically propose methods for automatically predicting the quality of articles.

Wikipedia has its definition of quality, too. For instance, the English Wikipedia content assessment guidelines [163] indicate that the most outstanding articles must be well-written, comprehensive, well-researched, and follow their style guidelines [165], which relate to the IQ properties defined by Lee et al. [80]. However, that definition may vary even within Wikipedia: according to Jemielniak and Wilamowski [68], not all language cultures share the same understanding of quality, which is a vital aspect to consider when designing a multilingual solution for quality assessment.

Overall, quality is a subjective property, so it is difficult to design an objective definition for it. However, there are certainly measurable characteristics that people often relate to outstanding quality, and we plan to determine them and their correlation with excellence in written documents, better understanding which are the most effective methods for predicting it.

3 METHODOLOGY

This systematic review aims to answer the following research questions:

RQ1. What are the most commonly used methods for the automatic quality assessment of Wikipedia articles?
RQ2. How can machine learning be best applied to predict article quality, and how do different approaches compare?
RQ3. What are the most common article features and quality metrics used to evaluate article quality in Wikipedia? How do these features compare, and how do they affect the performance of automatic assessment methods?
RQ4. Which common themes and gaps are there in the literature concerning this topic, and how can existing studies be improved to increase the adoption of automatic methods for the quality assessment of Wikipedia?

We guided our selection process by the PRISMA statement [107, 108], which defines a set of guidelines for conducting systematic literature reviews. Our selection included two main selection stages: **Database Querying** and **Citation Tracking**. Figure 1 outlines the selection process by detailing the number of publications included in each phase. Initially, we selected a set of records using research databases, and next, we conducted citation tracking on a subset of the initial selection.

3.1 Selection through Database Querying

Our initial selection stage comprises 4 phases (or steps): Identification, Screening, Eligibility, and Inclusion.

3.1.1 Identification. We considered three primary data sources: Google Scholar, ACM Digital Library, and Web of Science. In all of them, we retrieved all results containing "Wikipedia" and "quality" in its title. Searching in the title significantly reduces the number of results (for instance, reduces Google Scholar results by 99.97%), allowing the screening of all the retrieved results. We present the exact query and number of results for each database in Table 1.

Table 1. Search queries submitted in each database

Database	Query	Results
Google Scholar(1)	allintitle:“wikipedia” “quality”	212
ACM	[Title: wikipedia] AND [Title: quality] AND [E-Publication Date: (01/01/2001 TO 01/31/2023)]	71
Web of Science(1)	Wikipedia quality(2)	96

(1) We applied a date range of 2001-2023
(2) We applied this query on the title field

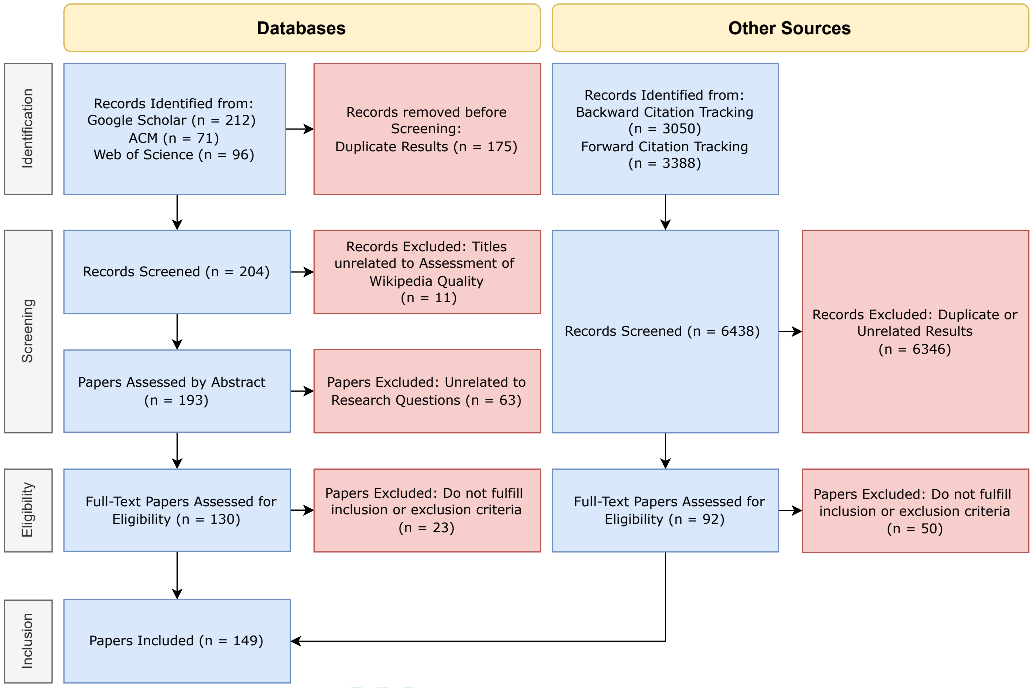
Databases
Other Sources
Identification
Screening
Eligibility
Inclusion
Records Identified from: Google Scholar (n = 212) ACM (n = 71) Web of Science (n = 96)
Records removed before Screening: Duplicate Results (n = 175)
Records Identified from: Backward Citation Tracking (n = 3050) Forward Citation Tracking (n = 3388)
Records Screened (n = 204)
Records Excluded: Titles unrelated to Assessment of Wikipedia Quality (n = 11)
Papers Assessed by Abstract (n = 193)
Papers Excluded: Unrelated to Research Questions (n = 63)
Records Screened (n = 6438)
Records Excluded: Duplicate or Unrelated Results (n = 6346)
Full-Text Papers Assessed for Eligibility (n = 130)
Papers Excluded: Do not fulfill inclusion or exclusion criteria (n = 23)
Full-Text Papers Assessed for Eligibility (n = 92)
Papers Excluded: Do not fulfill inclusion or exclusion criteria (n = 50)
Papers Included (n = 149)

Fig. 1. An overview of the selection process of our review

All database queries were run on January 24th, 2023, and were restricted to a date range of 2001-2023. We picked a lower limit of 2001 because Wikipedia was launched in that year [162], so we would unlikely find related articles from an earlier date.

Our Identification phase returned 379 results, as shown in Table 1. We discarded 175 records as duplicates.

3.1.2 Screening. Next, publication titles were assessed for possible relevance within the research area. All results that appeared at least marginally related to the quality assessment of Wikipedia proceeded to the next phase. Nearly all results moved forward, as our query was already somewhat strict. From the 204 non-duplicate titles, 193 were considered possibly relevant for our research. For instance, Salutari et al.'s study [123] was one of the results our query retrieved, but we deemed its title ("A Large-Scale Study of Wikipedia Users' Quality of Experience") unrelated to our research topic.

In this step, we also excluded results that did not respect the usual research article format. This includes theses, dissertations, and technical reports, among others. We opted to include pre-prints to avoid the exclusion of potentially insightful papers.

Finally, we scanned the papers' abstracts, focusing on the research questions. Only studies that propose automatic methods for measuring Wikipedia quality were included. This phase excluded publications experimenting with manual quality assessment approaches within specific sub-fields (e.g., Health) and studies whose abstracts we could not find. We advanced 130 studies to the next step.

3.1.3 Eligibility. In the Eligibility phase, we first defined clear inclusion and exclusion criteria and assessed all the manuscripts that reached this step. These criteria are described in Tables 2 and 3, and resulted in the exclusion of 23 results. We manually excluded all the publications that did not meet any inclusion criteria and publications that met at least one exclusion criterion.

Table 2. Inclusion Criteria

ID	Criteria
I1	Paper discusses machine learning approaches to predict information quality.
I2	Paper discusses possible features or metrics to assess information quality.

Table 3. Exclusion Criteria

ID	Criteria
E1	Paper does not discuss the assessment or prediction of the quality of a collaborative network. This includes studies that only discuss approaches to assess vandalism, controversy, or trust.
E2	Paper discusses manual approaches to assess article quality, as opposed to automatic ones.
E3	Paper is not in the English language.

3.1.4 Inclusion. In this phase, we run a serious analysis of each paper to collect all information relevant to our study, as we will describe in Section 3.3. Overall, this initial selection stage included 107 papers.

3.2 Selection through Citation Tracking

To minimize the probability of excluding relevant articles, we run citation tracking [51], searching through the references (backward tracking) and citations (forward tracking) of all included articles to identify potentially useful results.

Naturally, this procedure directly scales with the number of included articles and the respective number of references and citations. We determined that tracking the entire result set would be impractical, so we decided to only perform backward and forward tracking on the most relevant papers. We assessed relevance using a systematized scoring process, where we assign an integer value from 0 to 10 based on four questions, as listed in Table 4. We performed citation tracking on all results yielding a global relevance score of 4 or higher.

Table 4. Citation Tracking: Relevance Scoring Questions

ID	Question	Possible Scores*
Q1	Does the study focus on the topic of automatic quality assessment of Wikipedia?	From 0 (strongly disagree) to 3 (strongly agree)
Q2	Does the study describe and compare multiple ML approaches?	0 ($MLExp = 0$), 1 ($1 \leq MLExp < 4$), 2 ($4 \leq MLExp < 7$), and 3 ($MLExp \geq 7$)
Q3	Does the study describe and compare multiple article features and quality metrics?	0 ($FT = 0$), 1 ($1 \leq FT < 15$), 2 ($15 \leq FT < 50$), and 3 ($FT \geq 50$)
Q4	Does the study focus on an article language that's not English?	0 (No), 1 (Yes)

* MLExp corresponds to the number of used machine learning experiments and FT to the number of used features.

For each article, we manually checked the titles and abstracts of each reference and citation (we obtained citation data from Google Scholar in March of 2023), applying the same criteria used during the first Screening phase.

All relevant results transitioned to the Eligibility phase directly, therefore, will be assessed for inclusion and may end up being re-tracked, given a high enough relevance score. Overall, we performed this process on 92 different publications, which led to the inclusion of 42 new publications. Our systematic literature review included a total of 149 studies.

3.3 Data Collection

Throughout every phase of the selection process, we systematically logged all the data we collected and produced.

Initially, we store the title of every record we gathered during the first Identification phase and assign them a numeric identifier. We also store abstracts of publications that advanced to that sub-step of the Screening phase. Due to the substantial amount of inspected references and citations (6438), we did not keep any metadata for publications excluded during the Screening step of Citation Tracking.

We extracted most of the information during the Inclusion phase. We began by collecting relevant metadata of the 149 studies, such as the title, abstracts, keywords, authors, and year of publication. We then gathered study data, namely machine learning algorithms and respective performance, used article features and quality metrics, and dataset information.

All the information we collected is available in a research data repository[1], allowing readers to consult all the raw information we aggregated to display the results. We also provide a spreadsheet version of the dataset, similar to how we present it in this article, simplifying access for those who prefer not to handle the raw data directly.

4 OVERVIEW OF INCLUDED ARTICLES

This section provides an overview of the 149 included papers [3–6, 8, 9, 11–37, 39–44, 46–50, 52–64, 66–68, 70–72, 75, 77, 79, 81–100, 102–105, 109, 111–114, 117–122, 124, 126–146, 149–153, 155–159, 166, 168–174, 176–179, 181, 183, 184, 186], analyzing used methods and assessing metadata attributes, like publication venues, citation count, authors, and keywords.

4.1 Methods

Most papers (102 out of 149) follow one of these quality assessment strategies: classical learning (CL) models trained with article features, deep learning (DL) methods using full text or features, and metric-based approaches (MB). Many publications also study the correlation of specific features with quality: although they are not concrete automatic methods for quality prediction, we still consider them relevant for the purpose of this study. We summarize this information in Table 5.

4.1.1 Actionable Models and Visualization Tools. Designing an effective quality model for Wikipedia greatly assists Wikipedia users, by allowing easier identification of the best and worst articles. However, this does little for editors who wish to improve them. In the context of Explainable AI [175] it is important to create solutions that also suggest improvement paths, like the actionable model proposed by Warncke-Wang [159]. Some studies propose visualization tools that help solve this aspect. For instance, WikiRank [173] provides quality information and popularity stats of articles across many languages. Other studies [15, 26, 36] share a similar goal, although their solutions are much less thorough.

[1] While the paper awaits publication, the data is available at https://drive.google.com/drive/folders/1RWdsC79oQsTUfZtWrXu4IMv2WHHEOq44?usp=sharing

Table 5. Most popular approaches

Approach	# Papers	
CL + Features	51	[5, 6, 11, 12, 14, 19, 21–25, 27, 28, 34, 41–44, 46, 47, 49, 52, 53, 87, 91, 96, 99, 100, 105, 111, 113, 114, 118, 120, 122, 127, 131, 136, 137, 139, 150–152, 158, 159, 173, 174, 176–178, 183]
Deep Learning	20	[3, 9, 29–31, 50, 61, 92, 102, 109, 124, 126, 128–130, 155–157, 170, 184]
Metric-based	31	[18, 20, 33, 35, 39, 54, 56–58, 62–64, 67, 75, 84, 88, 93, 94, 104, 112, 119, 132, 140–145, 153, 172, 181]
Feat. Correlation	20	[32, 48, 55, 60, 66, 68, 70–72, 77, 81, 82, 86, 88, 90, 95, 103, 117, 133, 178]

4.1.2 Multi-language Assessment. As explained in Section 1, article quality varies significantly across different Wikipedia versions, so we tried to understand to what extent authors have studied quality assessment in multiple languages. Figure 2 shows that authors mostly focus on the English Wikipedia, but there are still some publications that consider other languages, occasionally within a machine learning context. We also discovered that 35 papers exclusively consider non-English Wikipedias [13, 15, 20, 26, 36, 39–41, 59, 60, 66, 71, 85, 89, 95, 120–122, 127, 131, 132, 140–146, 150, 171, 172, 174, 177, 178, 186].

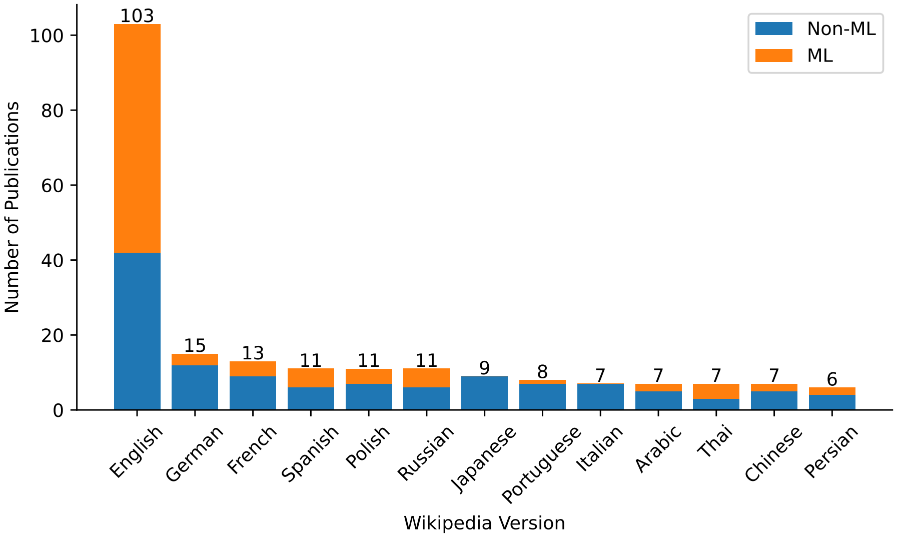

Fig. 2. Most commonly studied Wikipedia versions in included publications (only versions with more than 5 publications are shown)

We also analyzed how frequently authors study multiple versions at the same time. Figure 3 shows that papers almost never evaluate the quality of more than one language, but one of them [53] does a great job exploring this topic, designing different quality models for ten Wikipedia versions.

4.2 Year of Publication

We analyzed the publication year of our results to study trends in this topic. Stvilia et al.'s [135, 136] studies, from 2005, were the oldest of the 149, after which interest started to grow steadily. Figure 4 shows that classical learning

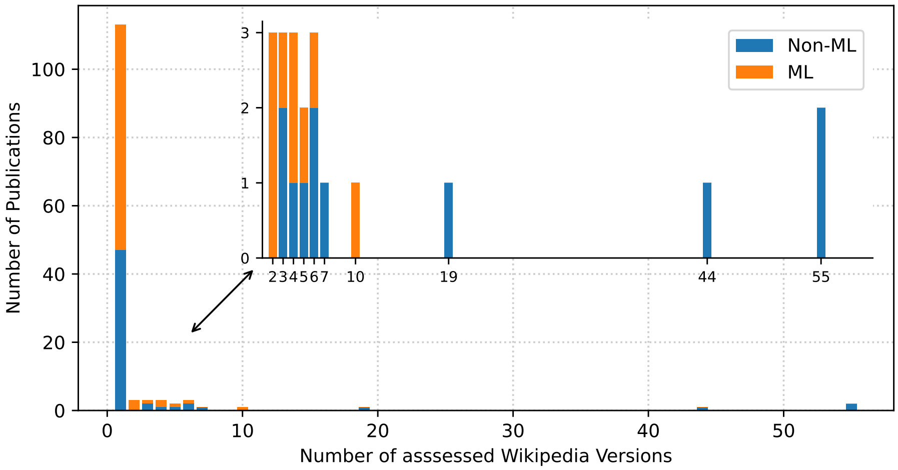

Fig. 3. Number of assessed Wikipedia Versions per publication

methods remained common through the years, but deep learning is clearly becoming a more prevalent approach, while metric-based studies are becoming more scarce.

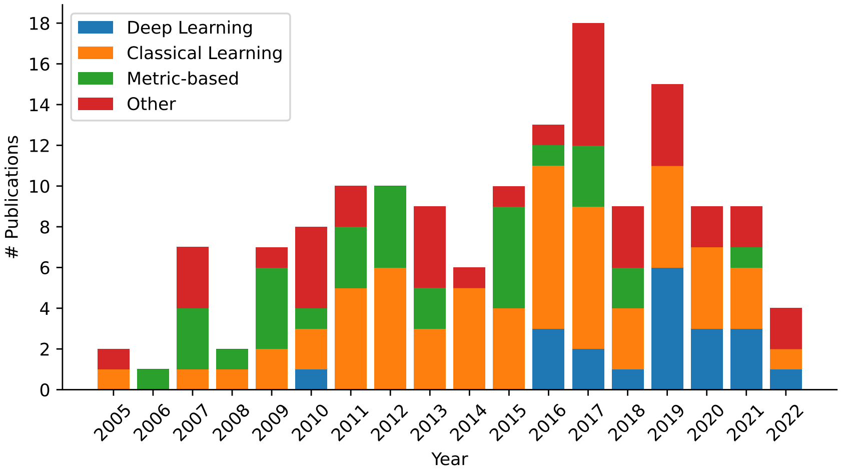

Fig. 4. Included papers by year of publication

4.3 Publication Venues

Most of the analyzed publications were published at international conferences, but we still counted many journal papers. To obtain a better overview of which publication venues are more frequent, we aggregated that information in Table 6, which shows the conferences and journals that published more than one of the papers we included in the review.

Table 6. Overview of the venues

Type	Venue	# Papers
Conference	OpenSym*: International Symposium on Open Collaboration	14
Conference	WWW: The Web Conference	6
Conference	BIS: International Conference on Business Information Systems	6
Conference	JCDL: ACM/IEEE Joint Conference on Digital Libraries	5
Journal	Journal of the Association for Information Science and Technology	5
Conference	ICIST: International Conference on Information and Software Technologies	4
Conference	CIKM: International Conference on Information and Knowledge Management	3
Journal	Proceedings of the ACM on Human-Computer Interaction	3
Conference	CLEF: Conference and Labs of the Evaluation Forum	2
Journal	Expert Systems with Applications	2
Journal	Online Information Review	2
Journal	Journal of Information Processing	2
Conference	WorldCIST: World Conference on Information Systems and Technologies	2
Conference	HT: ACM Conference on Hypertext & Social Media	2
Conference	WebMedia: Brazilian Symposium on Multimedia and the Web	2
Conference	CACIC: Argentine Congress of Computer Science	2
Conference	iSAI-NLP: International Joint Symposium on Artificial Intelligence and Natural Language Processing	2
Conference	WAIM: International Conference on Web-Age Information Management	2
Conference	WI: IEEE WIC ACM International Conference on Web Intelligence	2

* Formerly WikiSym

Notably, OpenSym [2] (formerly WikiSym) is the venue that has the most publications related to our research topic. That observation is not surprising considering their significant dedication to open collaboration research. Similarly, JASIST [3] is the peer-reviewed journal that publishes most articles on this topic.

4.4 Publication Influence - References and Citations

We examined and compared the number of citations and references of each included record, aiming to discover which papers were the most influential and which ones cover more sources. As shown in Figure 5, citation count varies significantly across the literature, but there are still many highly cited papers. The reference count is more stable, generally between 15 and 40. For legibility purposes, we excluded outliers (fliers) from the box plot, but we still find their analysis relevant. We found several highly cited papers, such as Stvilia et al.'s [134, 136], Wilkinson and Huberman's [166], Blumenstock's [14], and Hu et al.'s [62], all of which collect over 300 citations each. The publication with the most references is Halfaker and Geiger's [53], referencing 113 other papers.

We can obtain additional conclusions from Table 7, which shows that deep learning methods are not as influential as the others. However, as we have seen, these solutions are just starting to emerge, so it is possible this observation changes in the future.

[2] OpenSym Website URL: https://opensym.org/about-us/
[3] JASIST Website URL: https://asistdl.onlinelibrary.wiley.com/journal/23301643

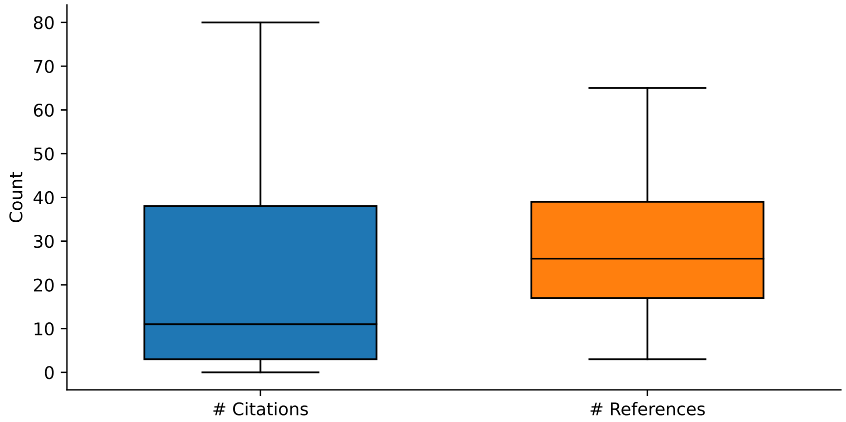

Fig. 5. Overview of citation (obtained from Google Scholar in March of 2023) and reference count of included publications

Table 7. Top 15 most impactful publications

Study	**Type**(1)	**Impact**(2)
A framework for information quality assessment [134]	FMC	37.88 / 13.12 / 19.19
Size matters: word count as a measure of quality on wikipedia [14]	CL	25.27 / N/A / 14.53
Cooperation and quality in wikipedia [166]	FMC	24.38 / N/A / 10.12
ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia [53]	CL	23.33 / N/A / 7.67
Measuring article quality in wikipedia: models and evaluation [62]	MB	22.5 / N/A / 11.38
Assessing information quality of a community-based encyclopedia [136]	CL	22.28 / N/A / 11.5
Who does what: Collaboration patterns in the wikipedia and their impact on article quality [97]	FMC	14.33 / N/A / 8.58
Tell me more: an actionable quality model for Wikipedia [159]	CL	12.6 / N/A / 8.1
NwQM: A Neural Quality Assessment Framework for Wikipedia [50]	DL	12.33 / 1.0 / 1.67
Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia [21]	CL	10.71 / 3.29 / 6.36
Assessing the quality of information on wikipedia: A deep-learning approach [156]	DL	10.67 / 3.67 / 5.67
Assessing the quality of Wikipedia articles with lifecycle based metrics [172]	MB	10.0 / N/A / 5.43
Predicting quality flaws in user-generated content: the case of wikipedia [6]	CL	9.91 / 2.82 / 5.18
Who Did What: Editor Role Identification in Wikipedia [179]	FMC	9.57 / N/A / 5.86
Information quality discussions in wikipedia [135]	FMC	9.33 / N/A / N/A

(1) CL = Classical Learning, DL = Deep Learning, MB = Metric-based, FMC = Feature/Metric quality correlation
(2) Measured as the number of citations per year of publication. Citation data obtained from Google Scholar, Web of Science, and Scopus, respectively, in March of 2023. N/A indicates the publication was not found in the respective database.

4.5 Abstract and Keywords Analysis

We also analyzed the most common terms in the abstract and keywords of the included publications. To do this, we first performed text normalization [4], which included tokenization, conversion to lowercase, removal of stop words and punctuation, and simplification of all words to their singular form. Next, for each group (abstract or keywords), we computed two measures: (1) the number of times each term appears in the collection of abstracts or keywords; (2)

[4] We used the NTLK library (https://www.nltk.org/) to assist in the normalization steps.

the number of abstracts or keywords in which each term appears in. These concepts are, respectively, the collection frequency (cf_t) and document frequency (df_t) as coined in the Information Retrieval area [101]. Table 8 summarizes this information, but most results are unsurprising. Aside from the obvious terms (e.g., quality, wikipedia, article), we can see that terms related to machine learning, edits, and network analysis frequently occur.

Table 8. Term Analysis: Ten highest Collection and Document Frequency of Abstract (cf_{ta}, df_{ta}) and Keywords (cf_{tk}, df_{tk}) terms

Term (t)	cf_{ta}	Term (t)	df_{ta}	Term (t)	cf_{tk}	Term (t)	df_{tk}
quality	652	quality	146	quality	116	wikipedia	104
article	594	wikipedia	139	wikipedia	106	quality	96
wikipedia	472	article	136	information	34	information	30
model	151	content	73	article	32	article	29
feature	146	paper	72	learning	22	assessment	20
content	133	result	65	assessment	21	learning	20
information	117	information	63	data	14	analysis	13
approach	98	model	58	analysis	13	network	13
editor	85	approach	54	network	13	classification	12
assessment	85	feature	54	edit	13	machine	12

4.6 Authors and Affiliations

We also decided to measure author presence across this research topic to determine which researchers study this subject more often. Table 9 summarizes this information, displaying all the authors from which we collected four or more publications, sorted by their influence, which is measured by averaging the number of citations per year of each paper we included.

Table 9. Authors with five or more included publications in the literature review

Author	Cits. / Year	Publications	(#)
Aaron L. Halfaker	9.73	[9, 52–54, 179]	(5)
Benno Stein	6.1	[4–6, 91, 96]	(5)
Quang-Vinh Dang	4.76	[27–31]	(5)
Pável Calado	4.37	[21–25, 100]	(6)
Daniel Hasan Dalip	3.82	[21–26, 100]	(7)
Marcos André Gonçalves	3.82	[21–26, 100]	(7)
Ping Wang	3.63	[61, 92, 155–157]	(5)
Witold Abramowicz	3.44	[84, 85, 87–90]	(6)
Marco Cristo	3.39	[21–25, 59, 60, 100]	(8)
Krzysztof Węcel	3.12	[71, 84–90, 173]	(9)
Włodzimierz Lewoniewski	2.88	[48, 71, 82–90, 173]	(12)
Yu Suzuki	1.64	[140–145]	(6)
Marcelo Errecalde	1.63	[36, 41–43, 91, 109, 111, 150, 153]	(9)
Edgardo Ferretti	1.62	[41–43, 91, 109, 111, 150]	(7)

4.7 Datasets, Source code, and External Tools

Regardless of the followed approach, authors generally create their datasets from Wikimedia dumps [5], selecting a subset of articles with varying quality distributions. Unfortunately, only 18 papers publish the datasets they use [11, 18, 23, 25, 28, 41, 43, 52, 62, 67, 79, 81, 109, 111, 126, 128, 184, 186], and most of the ones we encountered were inaccessible. In terms of implementation details, 10 papers provide the source code of their study [9, 18, 30, 31, 53, 70, 124, 129, 130, 186].

We also analyzed the used datasets to understand how they differ between studies. Figure 6 shows that machine learning datasets usually do not reach sizes as large as metric-based and feature-quality correlation approaches do. This observation makes sense: training models is computationally expensive, so studies that assess Wikipedia quality without artificial intelligence can afford to use larger datasets.

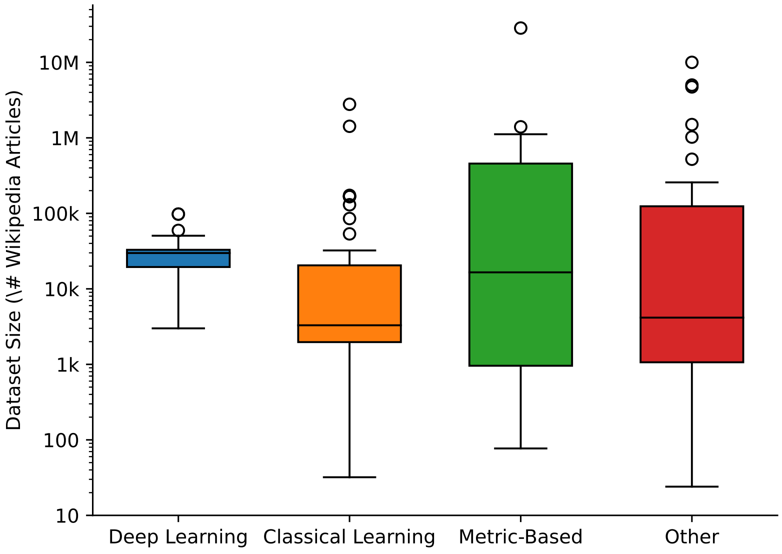

Fig. 6. Size of used datasets per method type

Finally, we analyzed the tools and libraries that authors most use in their studies. Table 10 shows some of the ones we collected when analyzing the manuscripts, which we hope will be useful for helping future researchers choose between technologies.

5 MACHINE LEARNING APPROACHES

Here we describe the approaches used by the 81 papers using machine learning to evaluate the quality of Wikipedia articles, comparing their performance. Authors do not always report their results using the same performance metrics, so it is not trivial to compare them directly. We wish to summarize the literature concisely but rigorously, so we will present this section's results sorted by performance value, clearly indicating the metric chosen by each study. We will not list here any of the 13 papers that do not use Accuracy, ROC AUC, or F1-score, but our dataset contains information related to every experiment.

[5] Dumps are available through https://dumps.wikimedia.org/

Table 10. Relevant Tools and Libraries

Type	Name	URL
NLP	Doc2Vec	https://radimrehurek.com/gensim/models/doc2vec.html
	Diction	https://www.gnu.org/software/diction/
	Snowball	https://snowballstem.org/
Classical Learning	Scikit-Learn	https://scikit-learn.org/
	Weka	https://waikato.github.io/weka-site/index.html
Deep Learning	Keras	https://keras.io/
	TensorFlow	https://www.tensorflow.org/
Wikipedia	MWParser	https://mwparserfromhell.readthedocs.io/en/latest/
	WebGraph	https://webgraph.di.unimi.it/

As Figure 7 suggests, 2-class and 6-class setups are much more common than the rest, so we will mainly focus on the performance of those solutions. We will separate those to allow us to better compare study performances, but we must first pay attention to the quality labels used for each class. Most 6-class studies consider Wikipedia's Stub to FA scale [163] (usually excluding A-tier), and 2-class typically follow a *Featured Article vs. Random Article* approach, so those comparisons should be safe. Also, since performance varies with both the number and distribution of considered classes, for each study we also show the dataset's imbalance ratio (*IR = # samples in the majority class / # samples in the minority class*) [185].

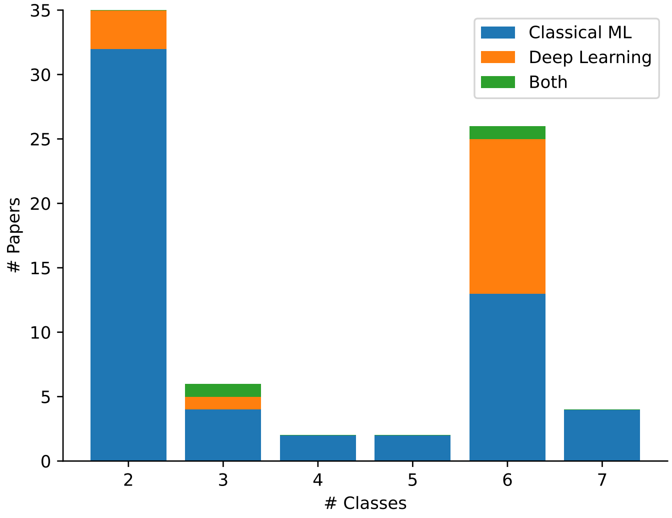

Fig. 7. Number of classes considered in machine learning experiments

Due to the nature of Machine Learning algorithms, it is unlikely that the best approach will be the same for every dataset. In fact, the *No Free Lunch* Theorem [167] states that all optimization algorithms have the same performance when averaged across all possible problems. Regardless, we collected all the results to understand better which algorithms were experimented with, and how performant they are in the given conditions, providing a baseline for future studies.

5.1 Classical Learning

We have seen that classical learning algorithms are the most common methods in this review: 65 publications opt to use them [4–6, 11–14, 16, 19, 21–25, 27, 28, 34, 41–44, 46, 47, 49, 52, 53, 85, 87, 89, 91, 96, 99, 100, 105, 109, 111, 113, 114, 118, 120–122, 126, 127, 131, 136–139, 146, 150–152, 155, 157–159, 168, 169, 173, 174, 176–178, 183]. Tables 11 and 12 show that decision trees, random forests, and SVMs are frequently great classical approaches, but the used performance metrics and class distribution vary so much that it is difficult to determine which solution is best. We also noticed that the best methods are almost always trained on English data, and those that are trained on multiple languages typically show much worse results on non-English data (e.g., Halfaker and Geiger [53]), which suggests there is a need for more work on multilingual assessment.

Table 11. Classical Learning accuracy of 6-class approaches

Study	Best Method	Accuracy	F1	AUC	IR*	Lang.
Włodzimierz et al. [87]	Random Forest	-	-	0.90	1.00	Multiple
Vittoria et al. [19]	Random Forest	-	-	0.89	95.06	English
Schmidt and Zangerle [126]	Gradient Boosted Trees	73.00%	-	-	1.10	English
Dang and Ignat [28]	Random Forest	64.00%	-	-	1.77	English
Halfaker [52]	ORES (Gradient Boosting)	62.90%	-	-	1.12	English
Halfaker and Geiger [53]	Gradient Boosting	62.90%	-	-	1.10	Multiple
Narun et al. [113]	MLR	49.35%	-	-	1.00	English

* Imbalance Ratio (IR) = # samples in the majority class / # samples in the minority class.

Table 12. Classical Learning accuracy of 2-class approaches (Top 10)

Study	Best Method	Accuracy	F1	AUC	IR*	Lang.
Saengthongpattana and Soonthornphisaj [120]	Naive Bayes	-	-	0.99	236.00	Thai
Blumenstock [14]	MLP	97.15%	-	-	6.12	English
Ofek and Rokach [105]	Bayes Network	-	-	0.97	1.00	English
Sugandhika and Ahangama [138]	Logistic Regression	96.00%	-	-	1.00	English
Adnan et al. [177]	Random Forest	95.50%	-	-	1.01	Arabic
Kui et al. [174]	C4.5 Decision Tree	94.60%	-	-	3.00	Chinese
Maik et al. [5]	Random Forest	-	-	0.94	1.00	English
Besiki et al. [136]	C4.5 Decision Tree	-	0.94	-	3.51	English
Lipka and Stein [96]	SVM	94.00%	-	-	1.00	English
Lian et al. [111]	SVM	-	0.94	-	1.00	English

* Imbalance Ratio (IR) = # samples in the majority class / # samples in the minority class.

Although quality might seem a continuous measure, almost all authors decided to solve a classification task. However, wrong predictions are typically not far from the correct ones. In fact, papers sometimes present off-by-one-class accuracy in their results [52], which tend to be much higher. Only eight studies [21–25, 27, 100, 127] tackle this problem as a regression task, but the method of solving it is very similar to others, typically using a feature-based approach.

5.2 Deep Learning

Although less common, deep learning methods have recently been gaining more relevance in this field. Among the 149 publications we collected, 20 of them use deep learning [3, 9, 29–31, 50, 61, 92, 102, 109, 124, 126, 128–130, 155–157, 170, 184]. Tables 13 and 14 suggest that LSTMs and GRUs lead to the most promising results, often better than classical methods. Unfortunately, we could not collect class distribution information from many studies, which makes us uncertain about how to best assess these results. Once again, we notice a strong preference for English datasets over non-English ones.

Table 13. Deep Learning accuracy of 6-class approaches

Study	Best Method	Accuracy	F1	AUC	IR*	Lang.
Jingrui et al. [61]	Stacked Learning	75.46%	-	-	?	English
Shiyue et al. [184]	RNN + LSTM	68.60%	-	-	1.16	English
Aili et al. [128]	Bi-LSTM+	68.17%	-	-	1.00	English
Dang and Ignat [31]	RNN + LSTM	68.00%	-	-	1.01	Multiple
Edison et al. [102]	Bi-LSTM	66.56%	-	-	?	Multiple
Bhanu et al. [50]	BERT + GRU	63.23%	-	-	1.64	English
Aili et al. [130]	BiLSTM	62.50%	-	-	1.02	English
Aili et al. [129]	BiLSTM	59.40%	-	-	?	English
Dang and Ignat [30]	DNN	55.00%	-	-	?	English
Dang and Ignat [29]	DNN	55.00%	-	-	?	English

* Imbalance Ratio (IR) = # samples in the majority class / # samples in the minority class. '?' indicates that we could not collect enough information about class distribution.

Table 14. Deep Learning accuracy of 2-class approaches

Study	Best Method	Accuracy	F1	AUC	IR*	Lang.
Muyan et al. [92]	BERT + GRU	97.58%	-	-	1.00	Multiple
Wang and Li [156]	Stacked LSTM	79.81%	-	-	?	English
Sumit et al. [9]	RNN	-	0.69	-	1.00	English

* Imbalance Ratio (IR) = # samples in the majority class / # samples in the minority class. '?' indicates that we could not collect enough information about class distribution.

6 ARTICLE FEATURES AND QUALITY METRICS

Some studies, typically deep learning ones, simply feed the article's full text to their model to obtain a quality prediction [29–31, 50, 61, 92, 124, 129, 130, 155], usually based on the Doc2Vec model [78]. However, most approaches still use article features or metrics, with and without machine learning.

This section overviews the article features and metrics we identified in this literature review. The distinction between features and metrics varies within the papers, sometimes used interchangeably. Here, we consider something a metric if it is not reasonably simple to compute and is used by the authors as a direct measure of quality (e.g., PeerReview [62]). In contrast, features are more straightforward and indirect quality measures (e.g., Character Count).

6.1 Article Features

We assigned a unique ID to all the features we collected from the reviewed publications, and each falls within one of the following categories:

- **Content features**, which relate to the length and structure of the article, taking into account factors such as the number of words, sections, or images.
- **Style features**, that measure how the authors write the articles, how long their phrases are, and what classes of words they use.
- **Readability features** estimate "the age or US grade level necessary to comprehend a text. (...) good articles should be well written, understandable, and free of unnecessary complexity" [21], by measuring the sentence and word complexity. They are characterized by their use of straightforward formulas that combine other types of features.
- **History features**, which analyze the review history of an article and related factors, namely the article's age and the number of contributions.
- **Network features** are a bit more complex, as they take into account the connections between Wikipedia articles to measure their influence.
- **Popularity features** track the engagement of the page, analyzing values related to the number of views and visitors.

We based this categorization on the work of previous authors (e.g., Bassani and Viviani [12], Dalip et al. [22]), but you may encounter slight modifications. For instance, we consider internal link counts as content features, as we believe that any measure that can be directly computed through an article's wikitext should belong to the Content, Style, or Readability category. Besides, although it is frequent for authors to assign internal link counts to the network category, external link counts are rarely considered network features, and we preferred to maintain consistency. Additionally, authors usually consider Content, Style, and Readability to be subcategories of *Text Features*. However, we distinguish them as different types in this review, aiming to reduce the disparity between the number of features per category.

Besides assigning a category, we also classify article features into two extra dimensions: **actionable** and **multilingual**. A feature is actionable if it can directly suggest how to improve the quality of the respective article, as proposed by Warncke-Wang et al. [159]. For instance, a low character count may indicate that expanding the article is beneficial for its quality. As for features that are technically manipulable but not in a relevant manner to the overall goal (e.g. revision count), we do not consider them actionable. The multilingual dimension answers the question: "Can this feature be applied to All, Most, or Some Wikipedia languages?" This is a relevant aspect when assessing, for example, readability features, whose formulas are often designed specifically for the English language [7]. The process of evaluating these two dimensions was conducted by the authors of this study independently, and discrepancies between assessments were later discussed until an agreement was reached.

Overall, we collected 321 distinct features throughout the 149 analyzed articles. Figure 8 better displays the proportion of features per category and how they correlate to the other dimensions too.

In this sub-section, we will overview every feature category, listing the 25% most used features from each one (but never less than 15). We finish this sub-section by summarizing our findings.

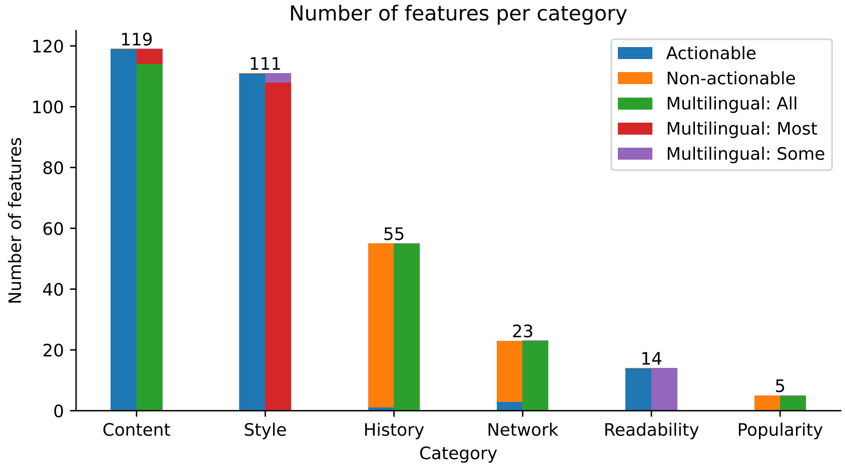

Fig. 8. Collected Features: Count per Category

6.1.1 *Content Features.*

Content features contain information regarding the length and structure of the article. They are one of the simplest to compute since they can be almost directly derived from the text. Table 15 lists the 25% most used content features gathered from the literature review.

Intuitively, there is a correlation between article length and quality. Good articles should not be too long and complex, overwhelming the reader, but not too simple either, as that could signify incomplete information. Also, according to Wikipedia [163], *stub* articles (drafts) are "usually very short", which also demonstrates that correlation.

Features related to the article structure are also essential to represent quality. Well-written articles should have a clear organization, with a balanced division of the content in sections and paragraphs. Images improve the reading experience, and references tend to increase the credibility of an article, so they are both decent indicators of quality.

6.1.2 *Style Features.*

Style features evaluate how contributors write their sentences, measuring the used word classes and sentence types. Table 16 provides the listing of the 25% most used style features. The intuition behind these features is more subtle, but the idea is that certain language practices are considered of better quality when presenting information about a topic [45].

6.1.3 *Readability Features.*

Readability features assess how easily readable a Wikipedia article is. They are calculated using carefully designed formulas that combine word, syllable, and character count and are listed in Table 17. We describe the formula for each one below, on Equations 1 to 11.

Automated Readability Index: Estimates readability by combining the average word length with the average sentence size.

$$ARI = 4.71\frac{characters}{words} + 0.5\frac{words}{sentences} - 21.43 \quad (1)$$

Table 15. List of 25% most used Content features mentioned in the assessed publications (complete feature set and paper citations in our dataset, file 'Full Feature List.pdf')

Name	Actionable	Multilingual	# Papers
Character Count (Article Length)	Yes	All	61
Internal Link Count	Yes	All	44
Image Count	Yes	All	42
External Link Count	Yes	All	35
Word Count	Yes	All	32
Section Count	Yes	All	32
Reference Count	Yes	All	32
Subsection Count	Yes	All	24
Sentence Count	Yes	All	22
Category Count	Yes	All	21
Citation Count	Yes	All	17
Information Noise Score (InfoNoise) (multiple definitions)	Yes	All	16
Longest Sentence Length (Words)	Yes	All	14
Has InfoBox	Yes	All	14
Subsection Count per Section (section nesting)	Yes	All	13
Mean Sentence Length (Words)	Yes	All	12
Paragraph Count	Yes	All	12
Image Count per Character	Yes	All	12
Image Count per Section	Yes	All	12
Heading Count	Yes	All	12
Mean Paragraph Length in Words	Yes	All	11
Reference Count per Character	Yes	All	11
Introduction Length (Lead section; abstract)	Yes	All	10
Mean Section Length in Words	Yes	All	9
Citation Count per Section	Yes	All	9
Table Count	Yes	All	9
Mean Word Length	Yes	All	8
Section Length Stdev.	Yes	All	8
Citation Count per Character	Yes	All	8

Coleman-Liau: Similarly to *ARI*, estimates readability by combining the average word length with the average sentence size.

$$CL = 5.88\frac{characters}{words} - 29.6\frac{sentences}{words} - 15.8 \tag{2}$$

Difficult Word Score (DWS): The DWS is calculated by counting the number of *difficult words*, which is a definition that varies between papers. According to Dang & Ignat, for example, "A word is considered difficult if it does not appear in a list of 3000 common English words that groups of fourth-grade American students could reliably understand." [28].

Dale-Chall: Also uses the concept of *difficult words*, combining it with the average sentence size to estimate readability.

$$DC = 0.1579 * (\frac{difficultwords}{words} * 100) + 0.0496\frac{words}{sentences} \tag{3}$$

Table 16. List of 25% most used Style features mentioned in the assessed publications (complete feature set and paper citations in our dataset, file 'Full Feature List.pdf')

Name	Actionable	Multilingual	# Papers
Short Sentence Rate (multiple definitions)	Yes	Most	14
Long Sentence Rate (multiple definitions)	Yes	Most	13
Passive Voice Sentence Count	Yes	Most	10
Question Count	Yes	Most	9
Auxiliary verb count	Yes	Most	9
Number of sentences starting with a pronoun	Yes	Most	8
To be Verb Count per Word	Yes	Most	8
Coordinate Conjunction Count per Word	Yes	Most	8
Number of sentences starting with an article	Yes	Most	7
Number of sentences starting with a coordinate conjunction	Yes	Most	7
Number of sentences starting with a subordinate preposition or conjunction	Yes	Most	7
Pronoun Count	Yes	Most	7
Preposition Count per Word	Yes	Most	7
Character/PoS N-grams	Yes	Some	7
Syllable Count	Yes	Most	6
Number of sentences starting with an interrogative pronoun	Yes	Most	6
Nominalization Count per Word	Yes	Most	6
Number of sentences starting with a preposition	Yes	Most	5
To be Verb Count	Yes	Most	5
Long Words Rate (multiple definitions)	Yes	Most	4
Question Count per Sentence	Yes	Most	4
Passive Voice Sentence Count per Sentence	Yes	Most	4
Syllable Count per Word	Yes	Most	4
One-Syllable Word Count	Yes	Most	4
Number of sentences starting with a pronoun per Sentence	Yes	Most	4
Number of sentences starting with an article per Sentence	Yes	Most	4

Flesch Reading Ease: Using the average sentence size and amount of syllables per word, computes a value between 0 and 100, where 0 indicates the text is difficult to understand.

$$FRE = 206.835 - 1.015\frac{words}{sentences} - 84.6\frac{syllables}{words} \tag{4}$$

Flesch-Kincaid Score: Same as *FRE*, but provides US grade levels instead of values between 0 and 100.

$$FK = 0.39\frac{words}{sentences} - 11.8\frac{syllables}{words} - 15.59 \tag{5}$$

FORCAST Readability Formula: Measures grade level from the number of monosyllabic words in a text sample of 150 words.

$$FOR = 20 - \frac{monosyllabic}{10} \tag{6}$$

Gunning Fog Index: Uses the concept of $complexwords$, which is the number of words with three or more syllables. The higher its value, the more difficult is the text to read.

$$GFI = 0.4(\frac{words}{sentences} + 100\frac{complexwords}{words}) \tag{7}$$

Lasbarhets Index (LIX): Very similar to GFI. In this case, $complexwords$, is the number of words with more than six characters. The higher its value, the more difficult is the text to read.

$$LIX = \frac{words}{sentences} + 100\frac{complexwords}{words} \tag{8}$$

Linsear Write Formula: Let n_1 be the number of words with two syllables or less, and n_2 be the number of words with three syllables or more.

$$LWF = \begin{cases} \frac{n_1+3\times n_2}{sentences\times 2}, & \text{if } \frac{n_1+3\times n_2}{sentences} > 20. \\ \frac{n_1+3\times n_2}{sentences\times 2} - 1, & \text{otherwise.} \end{cases} \tag{9}$$

Miyazaki Readability Score: Outputs a result between 0 and 100. The higher its value, the more difficult is the text to read.

$$MYZ = 164.935 - (18.792 * \frac{letters}{words} + 1.916\frac{words}{sentences}) \tag{10}$$

Smog-Grading: $polysyllables$ is the average number of polysyllabic words per 30 sentences (excluding proper names). They are usually calculated from a sample of 30 sentences.

$$SG = 3 + \sqrt{polysyllables} \tag{11}$$

Wiener Sachtextformel: The authors propose multiple formulas, but always aim to measure the grade level required to understand a German text [10].

6.1.4 History Features. History (or Review) features evaluate quality by analyzing an article's review history, authors, and contributions. The intuition behind them is that stable Wikipedia articles with trustworthy authors tend to be of better quality than controversial articles with very frequent edits. Table 18 lists the 25% most mentioned history features, some of which make use of the concepts of **Active review**: Review made by one of the most active 5% reviewers, and **Occasional review**: Review made by a user that edited the article less than four times.

6.1.5 Network features. Network features are based on the articles' graph, in which nodes represent articles and the edges show the links between them. High-quality articles tend to cite other sources and are more likely to be cited by others. Therefore, these features may be strong indicators of quality, and we list the 15 most used ones in Table 19.

6.1.6 Popularity Features. Popularity features are significantly less typical than the rest, as they require analytics information about the article. Still, ideally, high-quality articles would be more visited than worse ones. Table 20 lists all the popularity features we collected.

Table 17. List of all Readability features mentioned in the assessed publications (complete feature set and paper citations in our dataset, file 'Full Feature List.pdf')

Name	Actionable	Multilingual	# Papers
Flesch-Kincaid	Yes	Some	27
Flesch reading ease	Yes	Some	24
Automated Readability Index	Yes	Some	23
Coleman-Lieau	Yes	Some	20
Smog-Grading	Yes	Some	20
Gunning Fog Index	Yes	Some	19
Lasbarhets Index (LIX)	Yes	Some	13
Dale-Chall	Yes	Some	10
Linsear Write Formula	Yes	Some	6
Difficult Word Score	Yes	Some	5
Bormuth Readability Index	Yes	Some	4
Miyazaki Readability Score	Yes	Some	4
FORCAST Readability Formula	Yes	Some	3
Wiener Sachtextformel	Yes	Some	1

Table 18. List of 15 most used History features mentioned in the assessed publications (complete feature set and paper citations in our dataset, file 'Full Feature List.pdf')

Name	Actionable	Multilingual	# Papers
Revision Count	No	All	39
Contributor Count	No	All	39
Article Age (days)	No	All	34
Discussion Count	No	All	18
Anonymous Contributor Count	No	All	15
Revisions per Contributor	No	All	13
Revisions per Day	No	All	12
Registered Contributor Count	No	All	12
Registered Revision Count	No	All	12
Anonymous Revision Count	No	All	10
Current Revision Age (days) (currency)	No	All	9
Revisions per Contributor Stdev.	No	All	9
Recent Revision Count per Revision	No	All	9
Occasional Revision Count per Revision	No	All	8
Article Age per Revision	No	All	7

6.1.7 Summary. Most analyzed publications (104 out of 149) suggest or use article features for quality assessment. From those, we identified 1786 features, 321 unique. However, some papers do not always exhaustively detail the used features, so the real number may be much higher. For instance, Flekova et al. [46] claim to use 3,000, but we only managed to identify 25. Still, some authors focus much more on feature discovery than others. Bassani and Viviani [12], and Anderka et al. [6] both suggest more than 100 features in the ML solutions they propose, and ten papers suggest over 50 [6, 21, 22, 25, 25, 42, 100, 109, 156, 157].

We can also note a correlation between a feature's category and its suitability for multilingual or actionable models. With a few exceptions, that pattern generally falls into what is listed in Table 21.

Table 19. List of 15 most used Network features mentioned in the assessed publications (complete feature set and paper citations in our dataset, file 'Full Feature List.pdf')

Name	Actionable	Multilingual	# Papers
PageRank	No	All	20
Incoming Internal Link Count	No	All	16
In-degree	No	All	15
Out-degree	Yes	All	14
Local Clustering Coefficient	No	All	14
# of Versions in other Languages	Yes	All	14
Reciprocity within Neighbors	No	All	12
Assortativity (in-in, in-out, out-in, out-out)	No	All	9
Number of Articles with Common Editors (Connectivity)	No	All	4
Betweenness Centrality	No	All	3
Neighbor Count	Yes	All	2
Average Path Length	No	All	2
K-Core Number	No	All	2
HITS (Hyperlink-Induced Topic Selection)	No	All	2
Density	No	All	1

Table 20. List of all Popularity features mentioned in the assessed publications (complete feature set and paper citations in our dataset, file 'Full Feature List.pdf')

Name	Actionable	Multilingual	# Papers
Number of Page Visits	No	All	3
Number of Page Watchers	No	All	3
Number of Page Visits per Day	No	All	1
Number of Page Viewers	No	All	1
Number of Shares in Social Media	No	All	1

Table 21. Feature category vs Actionable and Multilingual properties

Category	Actionable	Multilingual
Content	Yes	All
Style	Yes	Most
Readability	Yes	Some
History	No	All
Network	No	All
Popularity	No	All

Finally, we decided to assess the contexts in which the collected features are used to create machine learning models. Figure 9 overviews the typical feature sets in those solutions, showing which categories are more common, and whether papers tend to use actionable and multilingual features. Content features are clearly the most prevalent, but otherwise, there is no strong preference for a specific category. We do see a preference for multilingual and actionable features in these models, which indicates that the existing literature may be useful to authors who wish to further explore these topics.

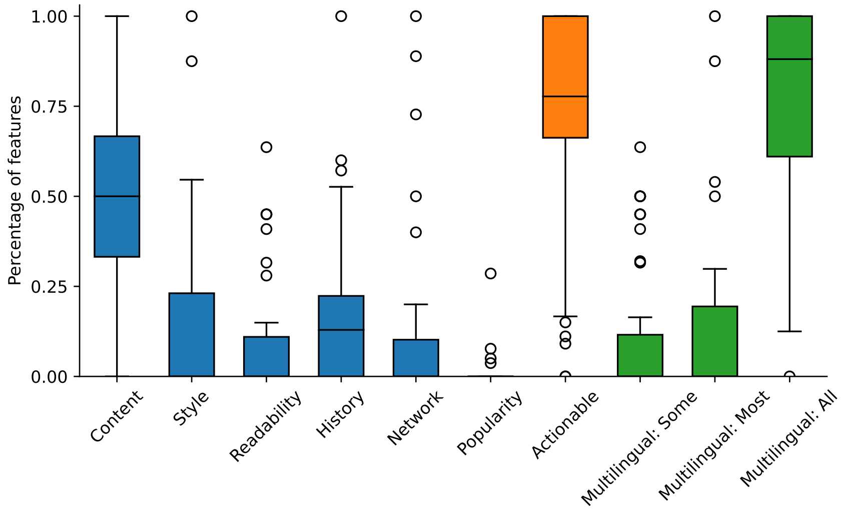

Fig. 9. Feature characteristics within machine learning approaches (Blue - Feature Category, Orange - Actionable, Green - Multilingual). The y-axis shows the percentage of the respective features among all features used in the paper.

6.2 Quality Metrics

Sometimes authors define new metrics, using them as direct measures of quality or as input for machine learning algorithms. We found that 31 of 149 included results use a direct metric-based approach, but 71 use metrics in their study, in some way. This section describes the two most common types of metrics within the literature: Stvilia's IQ metrics, and Text Survival metrics.

6.2.1 Stvilia's IQ Metrics. Stvilia et al. [134–136] propose 7 Information Quality (IQ) metrics which, combining different article features, aim to evaluate Wikipedia quality more systematically. We took the definitions for each metric directly from their study [134] and define their formulas in Equations 12 to 18.

Authority: Authority is defined as "the degree of the reputation of an information object in a given community". *Connectivity* corresponds to the number of articles with at least one contributor in common with the assessed article.

$$Authority = 0.2 \times UniqueEditors + 0.2 \times Contributions + 0.1 \times Connectivity + 0.3 \times Reverts + \\ 0.2 \times ExternalLinks + 0.1 \times RegisteredContributions + 0.2 \times AnonymousContributions \quad (12)$$

Completeness: Authors define Completeness as "the granularity or precision of an information object's model or content values".

$$Completeness = 0,4 \times InternalBrokenLinks + 0,4 \times InternalLinks + 0,2 \times ArticleLength \quad (13)$$

Complexity: Complexity is defined as “the degree of cognitive complexity of an information object relative to a particular activity”.

$$Complexity = 0,5 \times FleschreadingEase - 0,5 \times FleschKincaidgradelevel \tag{14}$$

Informativeness: Measures the amount of information in a document. $InfoNoise$ represents the ratio between the size of the information and the article, measuring the amount of *noise* in the document. *Diversity* refers to the ratio between editors and total edits.

$$Informativeness = 0,6 \times InfoNoise - 0,6 \times Diversity + 0,3 \times Images \tag{15}$$

Consistency: Consistency is defined as “the extent to which similar attributes or elements of an information object are consistently represented with the same structure, format, and precision”.

$$Consistency = 0,6 \times AdministratorsEditShare + 0,5 \times Age_{days} \tag{16}$$

Currency: Currency corresponds to “the age of an information object” in days.

$$Currency = CollectionDate - LastEditDate \tag{17}$$

Volatility: Volatility measures “the amount of time the information remains valid”.

$$Volatility = MedianRevertTime \tag{18}$$

Multiple authors talk about and experiment with these metrics [17, 18, 79, 138, 159], although rarely within a Machine Learning context. The application of Stvilia’s metrics to training ML algorithms could be worthy of experimentation.

6.2.2 Text Survival. Most metric-based approaches rely on the idea of *text survival*: If a piece of the article survives many revisions, that part is likely of good quality. The most common examples are *ProbReview* and *PeerReview*, initially proposed by Hu et al. [62, 63]. Many authors use *ProbReview* [12, 21–25, 100, 156, 157] and *PeerReview* [33, 144, 157] in their works, but *text survival* is a frequent concept in metric-based approaches. For example, the proposed measures of *Trensient Contribution* and *Persistent Contribution* [158, 172], *Word Persistence* [54], among others [34, 112, 140, 141, 143, 145] also rely on that notion.

7 DISCUSSION

This section discusses the results we obtained while attempting to answer the research questions we stated in Section 3.

7.1 RQ1. What are the most commonly used methods for the automatic quality assessment of Wikipedia articles?

We can identify three typical quality prediction strategies: metric-based approaches, classical machine learning models trained with article features, and deep learning methods using full text or features. Papers with other focus occasionally show up too. For example, Shen et al. [129, 130] proposed a multimodal classifier that uses both article features and a visual rendering of the document as input for quality prediction. There are also studies about quality flaw analysis and prediction [4–6, 41, 44, 92, 109, 150, 155], which identify frequent patterns of improvement.

We would also like to highlight Halfaker and Geiger's study [53], where they propose ORES: an API-based service that supports real-time scoring of Wikipedia edits, supporting many languages, and achieving exceptional results. The study has had a great impact in this research area and the service is currently provided by Wikimedia [6], setting a great benchmark for all future work.

It is challenging to determine the best strategies, as each one has its advantages and drawbacks. Metric-based approaches do not require model training, and some deep learning solutions are difficult to apply in a real-time scenario (e.g., Dang & Ignat's [31]). Regardless, our study showed that every strategy can perform effectively with the proper configuration (6-class: >60% accuracy, 2-class: >95% accuracy), and it is up to the researchers/developers to determine which solution makes the most sense for their context.

7.2 RQ2. How can machine learning be best applied to predict article quality, and how do different approaches compare?

Authors experiment with many machine learning algorithms, totaling 215 distinct experiments. It is not trivial to decide which algorithm is the most effective, as that is essentially dependent on the dataset definitions, but we see great performances from solutions using LSTMs, GRUs, Random Forests, and SVMs. Furthermore, boosting strategies, which combine multiple weak models into stronger ones [125] also tend to be very effective with classical algorithms (e.g., Decision Trees).

We have shown that most studies formulate this problem as a classification task, but we would like to note Teblunthuis's work [148], which shows that the English Wikipedia's quality levels (FA's, GA's, Stubs, etc.) are not evenly distributed on a linear scale, proposing a spacing that more effectively represents how distant are the multiple levels of quality. We did not include the study in this review, because it does not directly assess the quality of Wikipedia articles, but is still an insightful analysis of Wikipedia's quality scale.

Deep Learning is a topic that has gained significant popularity during this decade [180]. Most solutions we reviewed use classical approaches (e.g., decision trees, SVM), but we also discovered multiple deep learning solutions (e.g., LSTM). Additionally, we noticed that deep learning approaches were almost nonexistent ten years ago, while studies using classical methods for the automatic assessment of Wikipedia are not as common these days, relatively speaking.

Overall, although classical statistical learning approaches (e.g., decision trees, SVM) are more common, there has been a notable recent preference for deep learning (e.g., LSTM). Additionally, we noticed that deep learning approaches were almost nonexistent ten years ago, while studies using classical methods for the automatic assessment of Wikipedia are not as common nowadays, in relative terms. It is dangerous to directly compare results but, so far, deep learning seems to show slightly better performance than previous solutions. Furthermore, deep learning has been gaining significant

[6] ORES API: ttps://ores.wikimedia.org/

popularity during the past decade [180] and, if this trend continues, it is possible that their performance eclipses the effectiveness of classical algorithms.

7.3 RQ3. What are the most common article features and quality metrics used to evaluate article quality in Wikipedia? How do these features compare, and how do they affect the performance of automatic assessment methods?

In this review, we collected 321 different features, each of them factoring the text of an article, its review history, how it relates to other articles within Wikipedia, or even its popularity within users.

Even though Style features are the second largest group of identified features, they are one of the least used categories. Content features, which consider only the length and structure of the article, are both the most abundant and the most frequently used ones. The other features also appear often but not as much, possibly due to their higher computation complexity.

Not all papers use simple article features to assess quality, though. Some deep learning models train with the articles' full texts, and other studies opt for a metric-based approach, as shown in Section 5, but these approaches are not as common. Besides, although they may be used with Machine Learning, metrics are better suited for more manual approaches, so it is not surprising they do not show up as often in this review.

Some papers also compare different subsets of features regarding their effectiveness at predicting quality [21, 22, 25, 28, 44, 46, 156, 157]. By analyzing the different studies, we see that Content and Style features appear to be the most effective, but History and Network features are sometimes considered very useful for predicting quality too, so it would be wise to combine all categories. Readability Features also appear to generate decent results, but not so significantly as the others do, since those already combine existing features in a predefined way.

7.4 RQ4. Which common themes and gaps are there in the literature concerning this topic, and how can existing studies be improved to increase the adoption of automatic methods for the quality assessment of Wikipedia?

As the basis for a gap analysis, we organized multiple recurring methodology aspects into a frequency matrix, represented by the heatmap in Figure 10. Since we only pair the items two by two, this does not give us a picture of all the possible methods, but still provides quite some insight into what authors explore more and less within this field.

We can initially notice that the research within machine learning and feature-based models is extensive. There is also some work on metric-based solutions and studies focused on multiple Wikipedia languages. Nonetheless, there still exist some areas which the literature does not cover as much.

The most notable gap concerns the actionable solutions we previously discussed. Ideally, a model would not only predict article quality but also suggest possible steps for improvement. Warncke-Wang [159] proposes such a model, but the focus on actionable models seems to be otherwise scarce within the existing literature. Some studies [15, 26, 36, 173] propose reporting and visualization tools for analyzing the quality of Wikipedia, thus somewhat exploring this concept, but there is a clear lack of studies concerned about the topic. This review distinguishes actionable and non-actionable features in Section 6, aiming to guide authors in studying tools that assist Wikipedia readers and editors.

There is also very little work on multilingual solutions using machine learning, and none of those experiments with regression models. In addition, in studies that do design multilingual solutions, non-English model performance rarely comes close to the English one. For instance, the ORES study [53] predicts the quality of English articles with an accuracy of 62.9% but the French model's performance is only 44.2%. Wikipedia has millions of visitors using hundreds

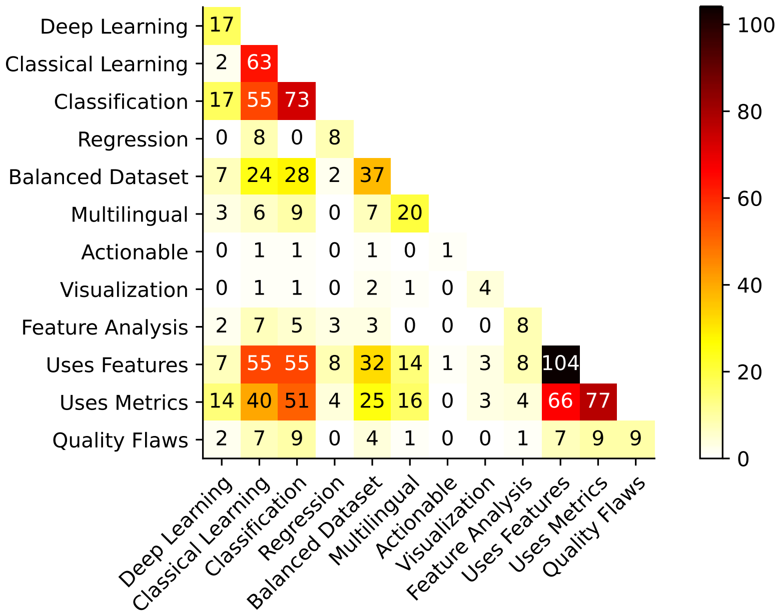

Fig. 10. Number of studies incorporating different methodological aspects

of versions [161], and as we discussed in Section 6, not every concept makes sense in every language, so it is crucial to study approaches that perform well in multiple languages.

There are some other issues not represented in the heatmap. For instance, there could also be a need for the existence of quality models specialized for specific Wikipedia categories. A couple of studies conduct several experiments with a few categories (e.g., history, biology, health), and some actually make models directed for specific fields [17, 46], but none do so with more than one area in the same study.

Finally, we were surprised not to see a lot of concern for the reproducibility and replicability [1] of the studies. Only ten papers share the source code of their work (from what we discovered, at least), and shared datasets are often inaccessible. We also sometimes struggled to locate even a description of the class distribution of the datasets, which is essential when comparing machine learning results. Hopefully, future authors will give increased importance to making their work more reproducible.

In summary, the quality assessment of Wikipedia is a significantly researched topic, for which there are many diverse methods to approach it. The results we collected will help any future work related to this field, and further experimentation may help develop better quality predictors.

8 CONCLUSIONS AND FUTURE WORK

This study reviewed literature related to the automatic assessment of the quality of Wikipedia articles, performing an in-depth analysis of 149 different papers out of thousands of inspected results. Our findings indicate that research on this topic has fluctuated for the past few years but only started getting attention several years after the launch of Wikipedia. There are many different proposals, but most use a feature-based traditional machine learning approach and refer to Wikipedia's content assessment standards to measure quality. We are starting to see more focus on deep learning methods, which may soon become the definite best option for this task, but it is difficult to compare results directly, since performance metrics, number of classes, and label distribution vary from study to study.

We can identify some limitations in our study, though. The most notable is the lack of non-bibliographic sources within our selection. Although our methodology should cover most journal submissions, conference papers, and other research repositories, some relevant studies may still be missed, such as Johnson's proposed quality model [69]. Nevertheless, nearly all relevant publications should be accessible through standard digital libraries.

Upon reviewing so much literature about the topic, we were puzzled by the fact that automatic assessment methods are still not widely used in Wikipedia. Although it is difficult to produce a direct answer, there are multiple potential explanations:

(1) Reliability: Even though some machine learning methods show impressive performance, model accuracies are far from 100%. This should not be a major issue, though, considering that, apart from B/C-tier articles, some papers show almost perfect one-off accuracy results.
(2) Complexity: As we have discussed before, quality is an extremely intricate concept with numerous properties, and some of them are much more challenging to assess than others. For instance, distinguishing a well-structured article from a poorly-structured one is trivial, compared to detecting false statements in a paragraph. Although this study does not focus so much on the trustworthiness part of information quality, all quality properties are relevant to Wikipedia users. As such, tools that do not fully grasp the essence of information quality may not be so well-received by the community.
(3) Accessibility: As we discussed in Section 7, there are not many reporting and visualization tools available for multilingual purposes, and we have seen that models are not easily transferable to other languages. We do have ORES [7] [53] and WikiRank [8] [173], but ORES is an API service directed to editors and WikiRank only provides a few actionable items for improvement. Besides, without a more direct integration with Wikipedia, it is difficult for a casual Wikipedia user to learn about those tools and know how to handle them.
(4) Self-regulation: Unlike most social media, Wikipedia has no central authority, and instead relies on collaborative moderation so, by design, it cannot have a ground-truth. This is why Wikimedia is reluctant to apply AI moderation to the website [147], and may also explain why automatic quality assessment methods are not as prevalent within Wikipedia.

We cannot solve all these impediments but we believe there is potential in combining existing approaches and making a tool accessible to every Wikipedia user, providing instantaneous feedback concerning the quality of the article. Such a project could promote the widespread usage of automatic Wikipedia quality models, and the results of this review are helpful indicators of which techniques lead to better performance. Still, future researchers must design and conduct their own set of experiments, comparing languages, features, metrics, and datasets, as model performance depends on much more than just its algorithm.

At the time of writing, OpenAI's GPT-4 has just been released [106], and the major tech companies are racing to compete against the novel ChatGPT [110]. Right now, the previously inconceivable idea of using a chatbot to predict Wikipedia article quality, explain its reasoning, and suggest items for improvement, sounds much more feasible. This

[7] ORES API: https://ores.wikimedia.org/
[8] WikiRank website: https://wikirank.net/

would require further research and development, and it is impossible to predict how this technology will advance in the near future, but it is evident how these tools could evolve to assist the topic of this research.

REFERENCES

[1] ACM. 2020. Artifact Review and Badging Version 1.1. https://www.acm.org/publications/policies/artifact-review-and-badging-current. Accessed: 2023-04-03.

[2] B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning Trust to Wikipedia Content. In *Proceedings of the 4th International Symposium on Wikis* (Porto, Portugal) *(WikiSym '08)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/1822258.1822293

[3] Rakshit Agrawal and Luca deAlfaro. 2016. Predicting the quality of user contributions via LSTMs. In *OpenSym '16: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–10. https://dl.acm.org/doi/10.1145/2957792.2957811

[4] Maik Anderka, Benno Stein, and Nedim Lipka. 2011. Detection of text quality flaws as a one-class classification problem. In *CIKM '11: ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York City, United States, 2313–2316. https://dl.acm.org/doi/10.1145/2063576.2063954

[5] Maik Anderka, Benno Stein, and Nedim Lipka. 2011. Towards automatic quality assurance in Wikipedia. In *WWW '11: International Conference Companion on World Wide Web*. Association for Computing Machinery, New York City, United States, 5–6. https://dl.acm.org/doi/10.1145/1963192.1963196

[6] Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting quality flaws in user-generated content: the case of wikipedia. In *SIGIR '12: International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York City, United States, 981–990. https://dl.acm.org/doi/10.1145/2348283.2348413

[7] Hélder Antunes and Carla Teixeira Lopes. 2019. Analyzing the Adequacy of Readability Indicators to a Non-English Language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11696 LNCS (2019), 149–155. https://doi.org/10.1007/978-3-030-28577-7_10/TABLES/3

[8] Ofer Arazy and Oded Nov. 2010. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *CSCW '10: Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, New York City, United States, 233–236. https://dl.acm.org/doi/10.1145/1718918.1718963

[9] Sumit Asthana, Sabrina Tobar Thommel, Aaron L. Halfaker, and Nikola Banovic. 2021. Automatically Labeling Low Quality Content on Wikipedia By Leveraging Patterns in Editing Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 23. Issue CSCW2. https://dl.acm.org/doi/10.1145/3479503

[10] Richard Bamberger and Annette T. Rabin. 1984. New Approaches to Readability: Austrian Research. *The Reading Teacher* 37, 6 (1984), 512–519. http://www.jstor.org/stable/20198517

[11] Elias Bassani and Marco Viviani. 2019. Automatically assessing the quality of Wikipedia contents. In *SAC '19: ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, New York City, United States, 804–807. https://dl.acm.org/doi/10.1145/3297280.3297357

[12] Elias Bassani and Marco Viviani. 2019. Quality of Wikipedia Articles: Analyzing Features and Building a Ground Truth for Supervised Classification. In *KDIR '19: International Conference on Knowledge Discovery and Information Retrieval*. Vienna University of Technology, Vienna, Austria, 338–346. https://www.scitepress.org/Link.aspx?doi=10.5220/0008149303380346

[13] Grace Gimon Betancourt, Armando Segnine, Carlos Trabuco, Amira Rezgui, and Nicolas Jullien. 2016. Mining team characteristics to predict Wikipedia article quality. In *OpenSym '16: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–9. https://dl.acm.org/doi/10.1145/2957792.2971802

[14] Joshua E. Blumenstock. 2008. Size matters: word count as a measure of quality on wikipedia. In *WWW '08: The Web Conference*. Association for Computing Machinery, New York City, United States, 1095–1096. https://dl.acm.org/doi/10.1145/1367497.1367673

[15] Fanny Chevalier, Stéphane Huot, and Jean-Daniel Fekete. 2010. WikipediaViz: Conveying article quality for casual Wikipedia readers. In *PacificVis '10: Pacific Visualization Symposium*. Institute of Electrical and Electronic Engineers, New York City, United States, 49–56. https://ieeexplore.ieee.org/document/5429611/

[16] Anamika Chhabra, Shubham Srivastava, S. R. S. Iyengar, and Poonam Saini. 2021. Structural Analysis of Wikigraph to Investigate Quality Grades of Wikipedia Articles. In *WWW '21: The Web Conference*. Association for Computing Machinery, New York City, United States, 584–590. https://dl.acm.org/doi/10.1145/3442442.3452345

[17] Luis Couto and Carla Teixeira Lopes. 2021. Assessing the quality of health-related Wikipedia articles with generic and specific metrics. In *WWW '21: The Web Conference*. Association for Computing Machinery, New York City, United States, 640–647. https://dl.acm.org/doi/10.1145/3442442.3452355

[18] Luis Couto and Carla Teixeira Lopes. 2021. Equal opportunities in the access to quality online health information? A multi-lingual study on Wikipedia. In *OpenSym '21: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–13. https://dl.acm.org/doi/10.1145/3479986.3480000

[19] Vittoria Cozza, Marinella Petrocchi, and Angelo Spognardi. 2016. A matter of words: NLP for quality evaluation of Wikipedia medical articles. In *IWCE '16: International Conference on Web Engineering*. Springer, Cham, Switzerland, 448–456. https://link.springer.com/chapter/10.1007/978-3-

319-38791-8_31

[20] Alberto Cusinato, Vincenzo Della Mea, Francesco Di Salvatore, and Stefano Mizzaro. 2009. QuWi: quality control in Wikipedia. In *WICOW '09: Workshop on Information Credibility on the Web*. Association for Computing Machinery, New York City, United States, 27–34. https://dl.acm.org/doi/10.1145/1526993.1527001

[21] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2009. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *JCDL '09: ACM/IEEE Joint Conference on Digital Libraries*. Association for Computing Machinery, New York City, United States, 295–304. https://dl.acm.org/doi/10.1145/1555400.1555449

[22] 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 (2011), 1–30. Issue 3. https://dl.acm.org/doi/10.1145/2063504.2063507

[23] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2012. On MultiView-Based Meta-learning for Automatic Quality Assessment of Wiki Articles. In *TPDL '12: International Conference on Theory and Practice of Digital Libraries*. Springer, Berlin, Heidelberg, 234–246. https://link.springer.com/chapter/10.1007/978-3-642-33290-6_26

[24] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2016. 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* 68 (2016), 286–308. Issue 2. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23650

[25] Daniel Hasan Dalip, Harlley Lima, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2014. Quality assessment of collaborative content with minimal information. In *JCDL '14: ACM/IEEE Joint Conference on Digital Libraries*. Association for Computing Machinery, New York City, United States, 201–210. https://dl.acm.org/doi/10.5555/2740769.2740804

[26] Daniel Hasan Dalip, Raquel Lara Santos, Diogo Rennó Rocha de Oliveira, Valéria Freitas Amaral, Marcos André Gonçalves, Raquel Oliveira Prates, Raquel C. M. Minardi, and Jussara Marques de Almeida. 2011. GreenWiki: a tool to support users' assessment of the quality of Wikipedia articles. In *JCDL '11: ACM/IEEE Joint Conference on Digital Libraries*. Association for Computing Machinery, New York City, United States, 469–470. https://dl.acm.org/doi/10.1145/1998076.1998190

[27] Quang-Vinh Dang. 2021. Assessing the Quality of Wikipedia Articles. In *ICMLSC '21: International Conference on Machine Learning and Soft Computing*. Association for Computing Machinery, New York City, United States, 1–4. https://dl.acm.org/doi/10.1145/3453800.3453801

[28] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2016. Measuring Quality of Collaboratively Edited Documents: The Case of Wikipedia. In *CIC '16: IEEE 2nd International Conference on Collaboration and Internet Computing*. Institute of Electrical and Electronic Engineers, New York City, United States, 266–275. https://ieeexplore.ieee.org/document/7809715

[29] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality assessment of wikipedia articles: a deep learning approach. https://dl.acm.org/doi/10.1145/2996442.2996447

[30] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality assessment of Wikipedia articles without feature engineering. In *JCDL '16: ACM/IEEE Joint Conference on Digital Libraries*. Association for Computing Machinery, New York City, United States, 27–30. https://dl.acm.org/doi/10.1145/2910896.2910917

[31] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2017. An end-to-end learning solution for assessing the quality of Wikipedia articles. In *OpenSym '17: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–10. https://dl.acm.org/doi/10.1145/3125433.3125448

[32] Paramita Das, Bhanu Prakash Reddy Guda, Sasi Bhusan Seelaboyina, Soumya Sarkar, and Animesh Mukherjee. 2021. Quality Change: Norm or Exception? Measurement, Analysis and Detection of Quality Change in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 6 (2021), 1 – 36. Issue CSCW1. https://dl.acm.org/doi/10.1145/3512959

[33] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2015. Measuring article quality in Wikipedia using the collaboration network. In *ASONAM '15: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Association for Computing Machinery, New York City, United States, 464–471. https://dl.acm.org/doi/10.1145/2808797.2808895

[34] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2017. *Structure-Based Features for Predicting the Quality of Articles in Wikipedia*. Springer, Cham, Switzerland. https://link.springer.com/chapter/10.1007/978-3-319-51049-1_6

[35] Huijing Deng, Bernadetta Tarigan, Mihai Grigore, and Juliana Sutanto. 2015. Understanding the 'Quality Motion' of Wikipedia Articles Through Semantic Convergence Analysis. In *HCIB '15: International Conference on HCI in Business*. Springer, Cham, Switzerland, 64–75. https://link.springer.com/chapter/10.1007/978-3-319-20895-4_7

[36] Cecilia di Sciascio, David Strohmaier, Marcelo Errecalde, and Eduardo Veas. 2017. WikiLyzer: Interactive Information Quality Assessment in Wikipedia. In *IUI '17: International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York City, United States, 377–388. https://dl.acm.org/doi/10.1145/3025171.3025201

[37] Pierpaolo Dondio and Stephen Barrett. 2007. Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project. *Informatica* 31 (2007), 151–160. Issue 2. https://arrow.tudublin.ie/scschcomart/25/

[38] Pierpaolo Dondio, Stephen Barrett, Stefan Weber, and Jean Marc Seigneur. 2006. Extracting Trust from Domain Analysis: A Case Study on the Wikipedia Project. In *Autonomic and Trusted Computing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 362–373.

[39] Gregory Druck, Gerome Miklau, and Andrew McCallum. 2008. Learning to Predict the Quality of Contributions to Wikipedia. https://maroo.cs.umass.edu/getpdf.php?id=834

[40] Fatemeh Fahimnia, Mansoureh Damerchiloo, Mohammad Khandan, and Mahshid Eltemasi. 2022. A Framework for Assessing the Quality of Wikipedia Articles: A Meta-synthesis of the Literature. *International Journal of Information Science and Management* 20 (2022), 91–118. Issue 1. https://www.magiran.com/paper/2379640

[41] Edgardo Ferretti, Leticia C. Cagnina, Viviana Paiz, Sebastián Delle Donne, Rodrigo Zacagnini, and Marcelo Errecalde. 2018. Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws. *Information Processing & Management* 54 (2018), 1169–1181. Issue 6. https://www.sciencedirect.com/science/article/pii/S0306457317309329?via%253Dihub

[42] Edgardo Ferretti, Donato Hernandez Fusilier, Rafael Guzmán-Cabrera, Manuel Montes y Gómez, Marcelo Errecalde, and Paolo Rosso. 2012. On the Use of PU Learning for Quality Flaw Prediction in Wikipedia. In *CLEF '12: Conference and Labs of the Evaluation Forum*. CLEF Initiative, Rome, Italy, 1178. https://www.researchgate.net/publication/236565329_On_the_Use_of_PU_Learning_for_Quality_Flaw_Prediction_in_Wikipedia

[43] Edgardo Ferretti, Matías Soria, Sebastián Pérez Casseignau, Lian Pohn, Guido Urquiza, Sergio Alejandro Gómez, and Marcelo Errecalde. 2017. Towards Information Quality Assurance in Spanish: Wikipedia. *Journal of Computer Science and Technology (JCS&T* 17 (2017), 29–36. Issue 1. https://www.semanticscholar.org/paper/8cba1878de84959de7a5401c9181819ee9bdf205

[44] Oliver Ferschke, Iryna Gurevych, and Marc Rittberger. 2012. FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia. In *CLEF '12: Conference and Labs of the Evaluation Forum*. CLEF Initiative, Rome, Italy, 1178. https://www.researchgate.net/publication/235982155_FlawFinder_A_Modular_System_for_Predicting_Quality_Flaws_in_Wikipedia

[45] Zeta Field. 2015. *How to write clearly*. Publications Office of the European Union, European Union. https://op.europa.eu/en/publication-detail/-/publication/725b7eb0-d92e-11e5-8fea-01aa75ed71a1

[46] Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. 2014. What makes a good biography?: multidimensional quality analysis based on wikipedia article feedback data. In *WWW '14: International Conference on World Wide Web*. Association for Computing Machinery, New York City, United States, 855–866. https://dl.acm.org/doi/10.1145/2566486.2567972

[47] Yasser Ganjisaffar, Sara Javanmardi, and Cristina Lopes. 2009. Review-based ranking of Wikipedia articles. In *CASON '09: International Conference on Computational Aspects of Social Networks*. Institute of Electrical and Electronic Engineers, New York City, United States, 98–104. https://ieeexplore.ieee.org/document/5176107/

[48] Mouzhi Ge and Włodzimierz Lewoniewski. 2020. Developing the Quality Model for Collaborative Open Data. *Procedia Computer Science* 176 (2020), 1883–1892. https://www.sciencedirect.com/science/article/pii/S187705092032130X

[49] Sindhuja Gopalan, Paolo Rosso, and Sobha Lalitha Devi. 2016. Discourse Connective - A Marker for Identifying Featured Articles in Biological Wikipedia. *Research in Computing Science* 117 (2016), 109–119. Issue 1. https://www.researchgate.net/journal/Research-in-Computing-Science-1870-4069

[50] Bhanu Prakash Reddy Guda, Sasi Bhusan Seelaboyina, Soumya Sarkar, and Animesh Mukherjee. 2020. NwQM: A Neural Quality Assessment Framework for Wikipedia. In *EMNLP '20: Conference on Empirical Methods in Natural Language Processing*. ACL Anthology, Online, 8396–8406. https://aclanthology.org/2020.emnlp-main.674/

[51] Neal R. Haddaway, Matthew J. Grainger, and Charles T. Gray. 2011. Citationchaser: A tool for transparent and efficient forward and backward citation chasing in systematic searching. *Research Synthesis Methods* 13 (2011), 533–545. Issue 4. https://doi.org/10.1002/jrsm.1563

[52] Aaron L. Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *OpenSym '17: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–9. https://dl.acm.org/doi/10.1145/3125433.3125475

[53] Aaron L. Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4 (2020), 1–37. Issue CSCW2. https://dl.acm.org/doi/10.1145/3415219

[54] Aaron L. Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A jury of your peers: quality, experience and ownership in Wikipedia. In *WikiSym '09: International Symposium on Wikis and Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–10. https://dl.acm.org/doi/10.1145/1641309.1641332

[55] Rainer Hammwöhner. 2010. Interlingual Aspects Of Wikipedia's Quality. https://epub.uni-regensburg.de/15572/

[56] Jingyu Han, Xiong Fu, Kejia Chen, and Chuandong Wang. 2011. Web Article Quality Assessment in Multi-dimensional Space. In *WAIM '11: International Conference on Web-Age Information Management*. Springer, Berlin, Heidelberg, 214–225. https://link.springer.com/chapter/10.1007/978-3-642-23535-1_20

[57] Jingyu Han, Chuandong Wang, Xiong Fu, and Kejia Chen. 2011. Probabilistic Quality Assessment of Articles Based on Learning Editing Patterns. In *CSSS '11: International Conference on Computer Science and Service System*. Institute of Electrical and Electronic Engineers, New York City, United States, 564–570. https://ieeexplore.ieee.org/abstract/document/5973947

[58] Jingyu Han, Chuandong Wang, and Dawei Jiang. 2011. Probabilistic Quality Assessment Based on Article's Revision History. In *DEXA '11: International Conference on Database and Expert Systems Applications*. Springer, Berlin, Heidelberg, 574–588. https://link.springer.com/chapter/10.1007/978-3-642-23091-2_50

[59] Raíza Hanada, Marco Cristo, and Maria da Graça Campos Pimentel. 2013. How do metrics of link analysis correlate to quality, relevance and popularity in wikipedia?. In *WebMedia '13: Brazilian Symposium on Multimedia and the Web*. Association for Computing Machinery, New York City, United States, 105–112. https://dl.acm.org/doi/10.1145/2526188.2526198

[60] Marcelo Yuji Himoro, Raíza Hanada, Marco Cristo, and Maria da Graça Campos Pimentel. 2013. An investigation of the relationship between the amount of extra-textual data and the quality of Wikipedia articles. In *WebMedia '13: Brazilian Symposium on Multimedia and the Web*. Association

for Computing Machinery, New York City, United States, 333–336. https://dl.acm.org/doi/10.1145/2526188.2526218

[61] Jingrui Hou, Jiangnan Li, and Ping Wang. 2021. Measuring Quality of Wikipedia Articles by Feature Fusion-based Stack Learning. In *ASIST '21: Association for Information Science and Technology*. Association for Information Science & Technology, Silver Spring, Maryland, 206–217. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/pra2.449

[62] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York City, United States, 243–252. https://dl.acm.org/doi/10.1145/1321440.1321476

[63] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. On improving wikipedia search using article quality. In *WIDM '07: ACM International Workshop on Web Information and Data Management*. Association for Computing Machinery, New York City, United States, 145–152. https://dl.acm.org/doi/10.1145/1316902.1316926

[64] Xiao Hu, Tzi-Dong Jeremy Ng, Lu Tian, and Chi-Un Lei. 2016. Automating assessment of collaborative writing quality in multiple stages: the case of wiki. In *LAK '16: International Conference on Learning Analytics & Knowledge*. Association for Computing Machinery, New York City, United States, 518–519. https://dl.acm.org/doi/abs/10.1145/2883851.2883963

[65] Christoph Hube and Besnik Fetahu. 2018. Detecting Biased Statements in Wikipedia. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1779–1786. https://doi.org/10.1145/3184558.3191640

[66] Myshkin Ingawale, Amitava Dutta, Rahul Roy, and Priya Seetharaman. 2013. Network analysis of user generated content quality in Wikipedia. *Online Information Review* 37 (2013), 602–619. Issue 4. https://www.emerald.com/insight/content/doi/10.1108/OIR-03-2011-0182/full/html

[67] Sara Javanmardi and Cristina Lopes. 2010. Statistical measure of quality in Wikipedia. In *SOMA '10: Workshop on Social Media Analytics*. Association for Computing Machinery, New York City, United States, 132–138. https://dl.acm.org/doi/10.1145/1964858.1964876

[68] Dariusz Jemielniak and Maciej Wilamowski. 2017. Cultural diversity of quality of information on Wikipedias. *Journal of the Association for Information Science and Technology* 68 (2017), 2460–2470. Issue 10. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23901

[69] Isaac Johnson. 2022. Language-agnostic Wikipedia article quality model card. https://meta.wikimedia.org/wiki/Machine_learning_models/Proposed/Language-agnostic_Wikipedia_article_quality_model_card. Accessed: 2023-04-04.

[70] Arash Joorabchi, Calibhe Doherty, and Jennifer Dawson. 2019. 'WP2Cochrane', a tool linking Wikipedia to the Cochrane Library: Results of a bibliometric analysis evaluating article quality and importance. *Health Informatics Journal* 26 (2019), 1881 – 1897. Issue 3. https://journals.sagepub.com/doi/10.1177/1460458219892711

[71] Nina Khairova, Włodzimierz Lewoniewski, and Krzysztof Węcel. 2017. Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. In *BIS '17: International Conference on Business Information Systems*. Springer, Cham, Switzerland, 28–40. https://link.springer.com/chapter/10.1007/978-3-319-59336-4_3

[72] Imran Khan, Shahid Hussain, Hina Gul, Muhammad Shahid, and Muhammad Jamal. 2019. An Empirical Study to Predict the Quality of Wikipedia Articles. In *WorldCIST '19: World Conference on Information Systems and Technologies*. Springer, Cham, Switzerland, 485–492. https://link.springer.com/chapter/10.1007/978-3-030-16187-3_47

[73] Khalid Al Khatib, Hinrich Schütze, and Cathleen Kantner. 2012. Automatic Detection of Point of View Differences in Wikipedia. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, 33–50. https://aclanthology.org/C12-1003

[74] Aniket Kittur, Bongwon Suh, and Ed H. Chi. 2008. Can You Ever Trust a Wiki? Impacting Perceived Trustworthiness in Wikipedia. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 477–480. https://doi.org/10.1145/1460563.1460639

[75] Rajmund Kleminski, Tomasz Kajdanowicz, Roman Bartusiak, and Przemyslaw Kazienko. 2017. On Quality Assesement in Wikipedia Articles Based on Markov Random Fields. In *ACIIDS '17: Asian Conference on Intelligent Information and Database Systems*. Springer, Cham, Switzerland, 782–791. https://link.springer.com/chapter/10.1007/978-3-319-54472-4_73

[76] Andrew Kuznetsov, Margeigh Novotny, Jessica Klein, Diego Saez-Trumper, and Aniket Kittur. 2022. Templates and Trust-o-Meters: Towards a Widely Deployable Indicator of Trust in Wikipedia. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 125, 17 pages. https://doi.org/10.1145/3491102.3517523

[77] Gabriel De la Calzada and Alex Dekhtyar. 2010. On measuring the quality of Wikipedia articles. In *WICOU '10: Workshop on Information Credibility on the Web*. Association for Computing Machinery, New York City, United States, 11–18. https://dl.acm.org/doi/10.1145/1772938.1772943

[78] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Beijing, China, 1188–1196. https://proceedings.mlr.press/v32/le14.html

[79] Tao-Chi Lee and Jayakrishnan Unnikrishnan. 2013. Monitoring network structure and content quality of signal processing articles on wikipedia. In *ICASSP '13: International Conference on Acoustics*. Institute of Electrical and Electronic Engineers, New York City, United States, 8766–8770. https://ieeexplore.ieee.org/document/6639378

[80] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang. 2002. AIMQ: a methodology for information quality assessment. *Information & Management* 40 (2002), 133–146. https://www.sciencedirect.com/science/article/abs/pii/S0378720602000435

[81] Jürgen Lerner and Alessandro Lomi. 2018. Knowledge categorization affects popularity and quality of Wikipedia articles. *PLoS ONE* 13 (2018), 1–22. Issue 1. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190674

[82] Włodzimierz Lewoniewski. 2017. Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis. In *BIS '17: International Conference on Business Information Systems*. Springer, Cham, Switzerland, 216–227. https://link.springer.com/chapter/10.1007/978-3-319-69023-0_19

[83] Włodzimierz Lewoniewski. 2018. Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia. In *BIS '18: International Conference on Business Information Systems*. Springer, Cham, Switzerland, 619–633. https://link.springer.com/chapter/10.1007/978-3-030-04849-5_53

[84] Włodzimierz Lewoniewski, Ralf-Christian Härting, Krzysztof Węcel, Christopher Reichstein, and Witold Abramowicz. 2018. Application of SEO Metrics to Determine the Quality of Wikipedia Articles and Their Sources. In *ICIST '18: International Conference on Information and Software Technologies*. Springer, Cham, Switzerland, 139–152. https://link.springer.com/chapter/10.1007/978-3-319-99972-2_11

[85] Włodzimierz Lewoniewski, Nina Khairova, Krzysztof Węcel, Nataliia Stratiienko, and Witold Abramowicz. 2017. Using Morphological and Semantic Features for the Quality Assessment of Russian Wikipedia. In *ICIST '17: International Conference on Information and Software Technologies*. Springer, Cham, Switzerland, 550–560. https://link.springer.com/chapter/10.1007/978-3-319-67642-5_46

[86] Włodzimierz Lewoniewski and Krzysztof Węcel. 2017. Relative Quality Assessment of Wikipedia Articles in Different Languages Using Synthetic Measure. In *BIS '17: International Conference on Business Information Systems*. Springer, Cham, Switzerland, 282–292. https://link.springer.com/chapter/10.1007/978-3-319-69023-0_24

[87] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2016. Quality and Importance of Wikipedia Articles in Different Languages. In *ICIST '16: International Conference on Information and Software Technologies*. Springer, Cham, Switzerland, 613–624. https://link.springer.com/chapter/10.1007/978-3-319-46254-7_50

[88] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics* 4 (2017), 43. Issue 4. https://www.mdpi.com/2227-9709/4/4/43

[89] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2018. Determining Quality of Articles in Polish Wikipedia Based on Linguistic Features. In *ICIST '18: International Conference on Information and Software Technologies*. Springer, Cham, Switzerland, 546–558. https://link.springer.com/chapter/10.1007/978-3-319-99972-2_45

[90] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2019. Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. *Computers* 8 (2019), 60. Issue 3. https://www.mdpi.com/2073-431X/8/3/60

[91] Elisabeth Lex, Michael Voelske, Marcelo Errecalde, Edgardo Ferretti, Leticia C. Cagnina, Christopher Horn, Benno Stein, and Michael Granitzer. 2012. Measuring the quality of web content using factual information. In *WebQuality '12: Joint WICOW/AIRWeb Workshop on Web Quality*. Association for Computing Machinery, New York City, United States, 7–10. https://dl.acm.org/doi/10.1145/2184305.2184308

[92] Muyan Li, Heshen Zhou, Jingrui Hou, Ping Wang, and Erpei Gao. 2022. Is cross-linguistic advert flaw detection in Wikipedia feasible? A multilingual-BERT-based transfer learning approach. *Knowledge-Based Systems* 252 (2022). Issue 109330. https://www.sciencedirect.com/science/article/pii/S0950705122006670

[93] Xinyi Li, Jintao Tang, Ting Wang, Zhunchen Luo, and Maarten de Rijke. 2015. Automatically Assessing Wikipedia Article Quality by Exploiting Article-Editor Networks. In *ECIR '15: European Conference on Information Retrieval*. Springer, Cham, Switzerland, 574–580. https://link.springer.com/chapter/10.1007/978-3-319-16354-3_64

[94] Ee-Peng Lim, Ba-Quy Vuong, Hady Wirawan Lauw, and Aixin Sun. 2006. Measuring Qualities of Articles Contributed by Online Communities. In *WI '16: IEEE WIC ACM International Conference on Web Intelligence*. Institute of Electrical and Electronic Engineers, New York City, United States, 81–87. https://ieeexplore.ieee.org/document/4061345

[95] Yan Lin and Chenxi Wang. 2020. Wisdom of crowds: the effect of participant composition and contribution behavior on Wikipedia article quality. *Journal of Knowledge Management* 24 (2020), 324–345. Issue 2. https://www.emerald.com/insight/content/doi/10.1108/JKM-08-2019-0416/full/html

[96] Nedim Lipka and Benno Stein. 2010. Identifying featured articles in wikipedia: writing style matters. In *WWW '10: International Conference on the World Wide Web*. Association for Computing Machinery, New York City, United States, 1147–1148. https://dl.acm.org/doi/10.1145/1772690.1772847

[97] Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems* 2 (2011), 1–23. Issue 2. https://dl.acm.org/doi/10.1145/1985347.1985352

[98] Jun Liu and Sudha Ram. 2018. Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering* 115 (2018), 80–93. https://www.sciencedirect.com/science/article/pii/S0169023X18300685?via%253Dihub

[99] Yuqing Lu, Lei Zhang, and Juan-Zi Li. 2013. Evaluating Article Quality and Editor Reputation in Wikipedia. In *CSWS '13: China Semantic Web Symposium and Web Science Conference*. Springer, Berlin, Heidelberg, 215–227. https://link.springer.com/chapter/10.1007/978-3-642-54025-7_19

[100] Luiz Felipe Gonçalves Magalhães, Marcos André Gonçalves, Sérgio Daniel Canuto, Daniel Hasan Dalip, Marco Cristo, and Pável Calado. 2019. Quality assessment of collaboratively-created web content with no manual intervention based on soft multi-view generation. *Expert Systems with Applications* 132 (2019), 226–238. https://www.sciencedirect.com/science/article/pii/S0957417419302830

[101] C.D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, England. https://books.google.pt/books?id=t1PoSh4uwVcC

[102] Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2019. An Edit-centric Approach for Wikipedia Article Quality Assessment. In *WNUT '19: Workshop on Noisy User-generated Text*. ACL Anthology, Online, 381–386. https://aclanthology.org/D19-5550/

[103] Emanuel Marzini, Angelo Spognardi, Ilaria Matteucci, Paolo Mori, Marinella Petrocchi, and Riccardo Conti. 2014. Improved Automatic Maturity Assessment of Wikipedia Medical Articles. In *OTM '14: Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg, 612–662. https://link.springer.com/chapter/10.1007/978-3-662-45563-0_37

[104] Sai T. Moturu and Huan Liu. 2009. Evaluating the trustworthiness of Wikipedia articles through quality and credibility. In *WikiSym '09: International Symposium on Wikis and Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–2. https://dl.acm.org/doi/10.1145/1641309.1641349

[105] Nir Ofek and Lior Rokach. 2015. A classifier to determine which Wikipedia biographies will be accepted. *Journal of the Association for Information Science and Technology* 66 (2015), 213–218. Issue 1. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23199

[106] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[107] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (3 2021). https://doi.org/10.1136/BMJ.N71

[108] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372 (2021), 1–36. https://doi.org/10.1136/bmj.n160 arXiv:https://www.bmj.com/content/372/bmj.n160.full.pdf

[109] Gerónimo Bazán Pereyra, Carolina Cuello, Gianfranco Capodici, Vanessa Jofré, Edgardo Ferretti, Rodolfo Bonnin, and Marcelo Errecalde. 2019. Predicting Information Quality Flaws in Wikipedia by Using Classical and Deep Learning Approaches. In *CACIC '19: Argentine Congress of Computer Science*. Springer, Cham, Switzerland, 3–18. https://link.springer.com/chapter/10.1007/978-3-030-48325-8_1

[110] David Pierce. 2023. ChatGPT started a new kind of AI race — and made text boxes cool again. https://www.theverge.com/2023/3/26/23655456/chatgpt-bard-bing-ai-race-text-boxes. Accessed: 2023-04-01.

[111] Lian Pohn, Edgardo Ferretti, and Marcelo Errecalde. 2014. Identifying featured articles in Spanish Wikipedia. http://sedici.unlp.edu.ar/bitstream/handle/10915/42288/Documento_completo.pdf?sequence=1

[112] Xiangju Qin and Pádraig Cunningham. 2012. Assessing the Quality of Wikipedia Pages Using Edit Longevity and Contributor Centrality. https://arxiv.org/abs/1206.2517

[113] Narun K. Raman, Nathaniel Sauerberg, Jonah Fisher, and Sneha Narayan. 2020. Classifying Wikipedia Article Quality With Revision History Networks. In *OpenSym '20: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–7. https://dl.acm.org/doi/10.1145/3412569.3412581

[114] Laura Rassbach, Trevor Blackford, and Brian Mingus. 2007. Exploring the Feasibility of Automatically Rating Online Article Quality. In *Wikimania '07: Wikimania Conference*. Wikimedia Foundation, San Francisco, California. https://scholar.google.pt/citations?view_op=view_citation%26hl=pt-PT%26user=T_sFnwoAAAAJ%26citation_for_view=T_sFnwoAAAAJ:u-x6o8ySG0sC

[115] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2020. A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2373–2380. https://aclanthology.org/2020.lrec-1.289

[116] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2022. Information asymmetry in Wikipedia across different languages: A statistical analysis. *Journal of the Association for Information Science and Technology* 73 (3 2022), 347–361. Issue 3.

[117] Thorsten Ruprechter, Tiago Santos, and Denis Helic. 2019. On the Relation of Edit Behavior, Link Structure, and Article Quality on Wikipedia. In *COMPLEX NETWORKS '19: International Workshop on Complex Networks & Their Applications*. Springer, Cham, Switzerland, 242–254. https://link.springer.com/chapter/10.1007/978-3-030-36683-4_20

[118] Thorsten Ruprechter, Tiago Santos, and Denis Helic. 2020. Relating Wikipedia article quality to edit behavior and link structure. *Applied Network Science* 5 (2020), 1–20. Issue 61. https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00305-y

[119] Giuseppe De Ruvo and Antonella Santone. 2015. Analysing wiki quality using probabilistic model checking. In *WET ICE '15: IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*. Institute of Electrical and Electronic Engineers, New York City, United States, 224–229. https://ieeexplore.ieee.org/document/71943655

[120] Kanchana Saengthongpattana and Nuanwan Soonthornphisaj. 2014. Assessing the Quality of Thai Wikipedia Articles Using Concept and Statistical Features. In *WorldCIST '14: World Conference on Information Systems and Technologies*. Springer, Cham, Switzerland, 513–523. https://link.springer.com/chapter/10.1007/978-3-319-05951-8_49

[121] Kanchana Saengthongpattana, Thepchai Supnithi, and Nuanwan Soonthornphisaj. 2017. Ontology-Based Classifiers for Wikipedia Article Quality Classification. In *iSAI-NLP '17: International Joint Symposium on Artificial Intelligence and Natural Language Processing*. Springer, Cham, Switzerland, 68–81. https://link.springer.com/chapter/10.1007/978-3-319-94703-7_7

[122] Kanchana Saengthongpattana, Thepchai Supnithi, and Nuanwan Soonthornphisaj. 2018. Quality Classification of ASEAN Wikipedia Articles using Statistical Features. In *iSAI-NLP '18: International Joint Symposium on Artificial Intelligence and Natural Language Processing*. Institute of Electrical and Electronic Engineers, New York City, United States, 1–6. https://ieeexplore.ieee.org/document/8692954/

[123] Flavia Salutari, Diego Da Hora Gilles, Dubuc, and Dario Rossi. 2019. A Large-Scale Study of Wikipedia Users' Quality of Experience. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3194–3200. https://doi.org/10.1145/3308558.3313467

[124] Soumya Sarkar, Bhanu Prakash Reddy Guda, Sandipan Sikdar, and Animesh Mukherjee. 2019. StRE: Self Attentive Edit Quality Prediction in Wikipedia. In *ACL '19: Annual Meeting of the Association for Computational Linguistics*. ACL Anthology, Online, 3962–3972. https://aclanthology.org/P19-1387/

[125] Robert E. Schapire. 2003. *The Boosting Approach to Machine Learning: An Overview*. Springer New York, New York, NY, 149–171. https://doi.org/10.1007/978-0-387-21579-2_9

[126] Manuel Schmidt and Eva Zangerle. 2019. Article quality classification on Wikipedia: introducing document embeddings and content features. In *OpenSym '19: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–8. https://dl.acm.org/doi/10.1145/3306446.3340831

[127] Seyedtaha Seyedsadr, Mohammadali Afsharkazemi, and Hashem Nikoomaram. 2016. Qualifying Articles of Persian Wikipedia Encyclopedia Through J48 Algorithm, ANFIS and Subtractive Clustering. *Automation* 3 (2016), 141–153. Issue 6. https://www.sciencepublishinggroup.com/journal/paperinfo?journalid=134%26doi=10.11648/j.acis.20150306.18

[128] Aili Shen, Jianzhong Qi, and Timothy Baldwin. 2017. A Hybrid Model for Quality Assessment of Wikipedia Articles. In *ALTA '17: Australasian Language Technology Association Workshop*. ACL Anthology, Online, 43–52. https://aclanthology.org/U17-1005/

[129] Aili Shen, Bahar Salehi, Timothy Baldwin, and Jianzhong Qi. 2019. A joint model for multimodal document quality assessment. In *JCDL '19: Joint Conference on Digital Libraries*. Association for Computing Machinery, New York City, United States, 107–110. https://dl.acm.org/doi/10.1109/JCDL.2019.00024

[130] Aili Shen, Bahar Salehi, Jainzhong Qi, and Timothy Baldwin. 2020. A multimodal approach to assessing document quality. *Journal of Artificial Intelligence Research* 68 (2020), 607–632. https://www.jair.org/index.php/jair/article/view/11647

[131] Nuanwan Soonthornphisaj and Peerapoom Paengporn. 2017. Thai Wikipedia article quality filtering algorithm. In *IMECS '17: International MultiConference of Engineers and Computer Scientists*. International Association of Engineers, Hong Kong, China. https://www.iaeng.org/publication/IMECS2017/IMECS2017_pp299-305.pdf

[132] Klaus Stein and Claudia Hess. 2007. Does it matter who contributes: a study on featured articles in the german wikipedia. In *HT '07: Conference on Hypertext and Hypermedia*. Association for Computing Machinery, New York City, United States, 171–174. https://dl.acm.org/doi/10.1145/1286240.1286290

[133] Besiki Stvilia, Abdullah Al-Faraj, and Yong Jeong Yi. 2009. Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research* 31 (2009), 232–239. Issue 4. https://www.sciencedirect.com/science/article/pii/S0740818809000954

[134] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. 2007. A framework for information quality assessment. *Journal of the Association for Information Science and Technology* 58 (2007), 1720–1733. Issue 12. https://onlinelibrary.wiley.com/doi/10.1002/asi.20652

[135] Besiki Stvilia, Michael B. Twidale, Les Gasser, and Linda C. Smith. 2005. Information quality discussions in wikipedia. In *ICKM '05: International Conference on Knowledge Management*. Universiti Putra Malaysia, Seri Kembangan, Malaysia. https://www.researchgate.net/publication/200773232_Information_Quality_Discussions_in_Wikipedia

[136] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. 2005. Assessing information quality of a community-based encyclopedia. In *ICIQ '05: International Conference on Information Quality*. Massachusetts Institute of Technology, Cambridge, Massachusetts, 442–454. https://www.semanticscholar.org/paper/Assessing-Information-Quality-of-a-Community-Based-Stvilia-Twidale/dd888dddccc2075a44f99ec2380fda652040afaf

[137] Qi Su and Pengyuan Liu. 2015. A Psycho-Lexical Approach to the Assessment of Information Quality on Wikipedia. In *WI-IAT '15: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Institute of Electrical and Electronic Engineers, New York City, United States, 184–187. https://ieeexplore.ieee.org/document/7397452

[138] Chinthani Sugandhika and Supunmali Ahangama. 2022. Assessing Information Quality of Wikipedia Articles through Google's E-A-T Model. *IEEE Access* 10 (2022), 52196–52209. https://ieeexplore.ieee.org/document/9770051

[139] Chinthani Sugandhika, Supunmali Ahangama, and Sapumal Ahangama. 2021. Modelling Wikipedia's Information Quality using Informativeness, Reliability and Authority. In *ICAC '21: International Conference on Advancements in Computing*. Institute of Electrical and Electronic Engineers, New York City, United States, 169–174. https://ieeexplore.ieee.org/document/9671092

[140] Yu Suzuki. 2012. Assessing Quality Values of Wikipedia Articles Using Implicit Positive and Negative Ratings. In *WAIM '12: International Conference on Web-Age Information Management*. Springer, Berlin, Heidelberg, 127–138. https://link.springer.com/chapter/10.1007/978-3-642-32281-5_13

[141] Yu Suzuki. 2013. Effects of Implicit Positive Ratings for Quality Assessment of Wikipedia Articles. *Journal of Information Processing* 21 (2013), 342–348. Issue 2. https://www.jstage.jst.go.jp/article/ipsjjip/21/2/21_342/_article

[142] Yu Suzuki. 2015. Quality Assessment of Wikipedia Articles Using h-index. *Journal of Information Processing* 23 (2015), 22–30. Issue 1. https://www.jstage.jst.go.jp/article/ipsjjip/23/1/23_22/_article

[143] Yu Suzuki and Masatoshi Yoshikawa. 2012. Mutual evaluation of editors and texts for assessing quality of Wikipedia articles. In *WikiSym '12: International Symposium on Wikis and Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–10. https://dl.acm.org/doi/10.1145/2462932.2462956

[144] Yu Suzuki and Masatoshi Yoshikawa. 2012. QualityRank: assessing quality of wikipedia articles by mutually evaluating editors and texts. In *HT '12: ACM Conference on Hypertext & Social Media*. Association for Computing Machinery, New York City, United States, 307–308. https://dl.acm.org/doi/10.1145/2309996.2310047

[145] Yu Suzuki and Masatoshi Yoshikawa. 2013. Assessing quality score of Wikipedia article using mutual evaluation of editors and texts. In *CIKM '13: ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York City, United States, 1722–1732. https://dl.acm.org/doi/10.1145/2505515.2505610

[146] Marcin Sydow, Katarzyna Baraniak, and Paweł Teisseyre. 2017. Diversity of editors and teams versus quality of cooperative work: experiments on wikipedia. *Journal of Intelligent Information Systems* 48 (2017), 601–632. https://link.springer.com/article/10.1007/s10844-016-0428-1

[147] Diego Sáez-Trumper. 2021. Disinformation and AI: The Differences Between Wikipedia and Social Media. https://diff.wikimedia.org/2021/09/15/disinformation-and-ai-the-differences-between-wikipedia-and-social-media/. Accessed: 2023-04-04.

[148] Nathan Teblunthuis. 2021. Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression. In *Proceedings of the 17th International Symposium on Open Collaboration* (Online, Spain) *(OpenSym '21)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. https://doi.org/10.1145/3479986.3479991

[149] Michail Tsikerdekis. 2017. Cumulative Experience and Recent Behavior and their Relation to Content Quality on Wikipedia. *Interacting with Computers* 29 (2017), 737–754. Issue 5. https://academic.oup.com/iwc/article/29/5/737/3885842

[150] Guido Urquiza, Matías Soria, Sebastián Pérez Casseignau, Edgardo Ferretti, Sergio Alejandro Gómez, and Marcelo Errecalde. 2016. On the Assessment of Information Quality in Spanish Wikipedia. In *CACIC '19: Argentine Congress of Computer Science*. National University of La Plata, La Plata, Argentina, 702–711. http://sedici.unlp.edu.ar/handle/10915/56750

[151] Srikar Velichety. 2019. Quality Assessment of Peer-Produced Content in Knowledge Repositories Using Big Data and Social Networks: The Case of Implicit Collaboration in Wikipedia. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems* 50 (2019), 28–51. Issue 4. https://dl.acm.org/doi/10.1145/3371041.3371045

[152] Srikar Velichety, Sudha Ram, and Jesse Bockstedt. 2019. Quality Assessment of Peer-Produced Content in Knowledge Repositories using Development and Coordination Activities. *Journal of Management Information Systems* 36 (2019), 478–512. Issue 2. https://www.tandfonline.com/doi/full/10.1080/07421222.2019.1598692

[153] Carlos G. Velázquez, Leticia C. Cagnina, and Marcelo Errecalde. 2017. On the Feasibility of External Factual Support as Wikipedia's Quality Metric. *Processamiento del Lenguaje Natural* 58 (2017), 93–100. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5417

[154] Nicholas Vincent and Brent Hecht. 2021. A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 18. Issue CSCW1. https://doi.org/10.1145/3449078

[155] Ping Wang, Muyan Li, Xiaodan Li, Heshen Zhou, and Jingrui Hou. 2021. A hybrid approach to classifying Wikipedia article quality flaws with feature fusion framework. *Expert Systems with Applications* 181 (2021), 115089. Issue 1. https://www.sciencedirect.com/science/article/pii/S0957417421005303?via%253Dihub

[156] Ping Wang and Xiaodan Li. 2020. Assessing the quality of information on wikipedia: A deep-learning approach. *Journal of the Association for Information Science and Technology* 71 (2020), 16–28. Issue 1. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24210

[157] Ping Wang, Xiaodan Li, and Renli Wu. 2019. A deep learning-based quality assessment model of collaboratively edited documents: A case study of Wikipedia. *Journal of Information Science* 47 (2019), 176 – 191. Issue 2. https://journals.sagepub.com/doi/10.1177/0165551519877646

[158] Se Wang and Mizuho Iwaihara. 2010. Quality Evaluation of Wikipedia Articles through Edit History and Editor Groups. In *APWeb '11: Asia-Pacific Web Conference*. Springer, Berlin, Heidelberg, 188–199. https://link.springer.com/chapter/10.1007/978-3-642-20291-9_20

[159] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell me more: an actionable quality model for Wikipedia. In *WikiSym '13: International Symposium on Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–10. https://dl.acm.org/doi/10.1145/2491055.2491063

[160] Wikimedia. 2022. Wikistats - Statistics For Wikimedia Projects. https://stats.wikimedia.org. Accessed: 2023-04-03.

[161] Wikipedia. 2022. List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed: 2023-04-03.

[162] Wikipedia. 2022. Wikipedia. https://en.wikipedia.org/wiki/Wikipedia. Accessed: 2023-04-03.

[163] Wikipedia. 2022. Wikipedia: Content assessment. https://en.wikipedia.org/wiki/Wikipedia:Content_assessment. Accessed: 2023-04-03.

[164] Wikipedia. 2022. Wikipedia: Size of Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. Accessed: 2023-04-03.

[165] Wikipedia. 2023. Wikipedia: Manual of Style. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/. Accessed: 2023-04-04.

[166] Dennis Wilkinson and Bernardo Huberman. 2007. Cooperation and quality in wikipedia. In *WikiSym '07: International Symposium on Wikis*. Association for Computing Machinery, New York City, United States, 157–164. https://dl.acm.org/doi/10.1145/1296951.1296968

[167] David H Wolpert and William G Macready. 1997. No Free Lunch Theorems for Optimization. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION* 1 (1997), 67. Issue 1.

[168] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. 2011. Characterizing Wikipedia pages using edit network motif profiles. In *SMUC '11: International Workshop on Search and Mining User-generated Contents*. Association for Computing Machinery, New York City, United States, 45–52. https://dl.acm.org/doi/10.1145/2065023.2065036

[169] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. 2012. Classifying Wikipedia articles using network motif counts and ratios. In *WikiSym '12: International Symposium on Wikis and Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–12. https://dl.acm.org/doi/10.1145/2462932.2462948

[170] Kewen Wu, Qinghua Zhu, Yuxiang Zhao, and Hua Zheng. 2010. Mining the Factors Affecting the Quality of Wikipedia Articles. In *ISME '10: International Conference of Information Science and Management Engineering*. Institute of Electrical and Electronic Engineers, New York City, United States, 343–346. https://ieeexplore.ieee.org/document/5572324

[171] Thomas Wöhner, Sebastian Köhler, and Ralf Peters. 2015. Good Authors = Good Articles? - How Wikis Work. In *WI '15: International Conference on Wirtschaftsinformatik*. Association for Information Systems, Atlanta, Georgia. https://aisel.aisnet.org/wi2015/59/?utm_source=aisel.aisnet.org%252Fwi2015%252F59%26utm_medium=PDF%26utm_campaign=PDFCoverPages

[172] Thomas Wöhner and Ralf Peters. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *WikiSym '09: International Symposium on Wikis and Open Collaboration*. Association for Computing Machinery, New York City, United States, 1–10. https://dl.acm.org/doi/10.1145/1641309.1641333

[173] Krzysztof Węcel and Włodzimierz Lewoniewski. 2015. Modelling the Quality of Attributes in Wikipedia Infoboxes. In *BIS '15: International Conference on Business Information Systems*. Springer, Cham, Switzerland, 308–320. https://link.springer.com/chapter/10.1007/978-3-319-26762-3_27

[174] Kui Xiao, Bing Li, Peng He, and Xi hui Yang. 2013. Detection of Article Qualities in the Chinese Wikipedia Based on C4.5 Decision Tree. In *KSEM '13: International Conference on Knowledge Science*. Springer, Berlin, Heidelberg, 444–452. https://link.springer.com/chapter/10.1007/978-3-642-39787-5_36

[175] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer International Publishing, Cham, 563–574.

[176] Yanxiang Xu and Tiejian Luo. 2011. Measuring article quality in Wikipedia: Lexical clue model. In *SWS '11: Symposium on Web Society*. Institute of Electrical and Electronic Engineers, New York City, United States, 141–146. https://ieeexplore.ieee.org/document/6101286

[177] Adnan Yahya, Afnan Ahmad, Alaa Assaf, Rawan Khater, and Ali Salhi. 2020. Models for Arabic Document Quality Assessment. In *BIS '20: International Conference on Business Information Systems*. Springer, Cham, Switzerland, 297–310. https://link.springer.com/chapter/10.1007/978-3-030-61146-0_24

[178] Adnan Yahya and Ali Salhi. 2014. Quality assessment of Arabic web content: The case of the Arabic Wikipedia. In *IIT '14: International Conference on Innovations in Information Technology*. Institute of Electrical and Electronic Engineers, New York City, United States, 36–41. https://ieeexplore.ieee.org/document/6987558

[179] Diyi Yang, Aaron L. Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia. In *ICWSM '16: International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence, Palo Alto, California, 446–455. https://ojs.aaai.org/index.php/ICWSM/article/view/14732

[180] M. Mutlu Yapıcı, Adem Tekerek, and Nurettin Topaloğlu. 2019. Literature Review of Deep Learning Research Areas. *Gazi Mühendislik Bilimleri Dergisi* 5 (2019), 188 – 215. Issue 3. https://doi.org/10.30855/gmbd.2019.03.01

[181] Linfeng Yu and Mizuho Iwaihara. 2018. Finding high quality documents through link and click graphs. In *IIAI-AAI '18: International Congress on Advanced Applied Informatics*. Institute of Electrical and Electronic Engineers, New York City, United States, 49–54. https://ieeexplore.ieee.org/abstract/document/8693372

[182] Honglei Zeng, Maher A Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. 2006. *Computing trust from revision history.* Technical Report. Stanford Univ Ca Knowledge Systems LAB. https://apps.dtic.mil/sti/citations/ADA454704

[183] Ning Zhang, Lingyun Ruan, and Luo Si. 2015. *Predicting Low-Quality Wikipedia Articles Using User's Judgements*. Springer, Cham, Switzerland. https://link.springer.com/chapter/10.1007/978-3-319-05467-4_6

[184] Shiyue Zhang, Zheng Hu, Chunhong Zhang, and Ke Yu. 2018. History-Based Article Quality Assessment on Wikipedia. In *BIGCOMP '18: International Conference on Big Data and Smart Computing*. Institute of Electrical and Electronic Engineers, New York City, United States, 1–8. https://ieeexplore.ieee.org/document/8367090

[185] Rui Zhu, Yiwen Guo, and Jing-Hao Xue. 2020. Adjusting the imbalance ratio by the dimensionality of imbalanced data. *Pattern Recognition Letters* 133 (2020), 217–223. https://doi.org/10.1016/j.patrec.2020.03.004

[186] Didem Ölçer and Tuğba Taşkaya Temizel. 2022. Quality assessment of web-based information on type 2 diabetes. *Online Information Review* 46 (2022), 715–732. Issue 4. https://www.emerald.com/insight/content/doi/10.1108/OIR-02-2021-0089/full/html