Włodzimierz Lewoniewski, PhD
specializes in studying information quality in open knowledge bases (such as Wikipedia, DBpedia, Wikidata), fake news detection, natural language processing, and machine learning.
wlodzimierz.lewoniewski@ue.poznan.pl

Information Quality on Wikipedia

Wikipedia, one of the world's most popular websites, owes its success to its authors – i.e. to all of us. But how do we know if the information it offers is reliable?

Włodzimierz Lewoniewski

Department of Information Systems
Poznań University of Economics and Business

The unrestricted nature of the Internet makes it possible to be exposed to a wide variety of viewpoints and opinions. There are, at present, over a billion websites offering information on various topics. Over many years, Wikipedia – a publicly accessible and editable encyclopedia – has risen to become one of the most well-known online sources of information. It currently contains over 60 million articles in more than 300 languages, making knowledge accessible to people from diverse cultures and regions. Additionally, many popular websites and tools (including Google's search engine and ChatGPT) currently utilize content from Wikipedia to improve the quality of their services.

A comparison of the different names used for quality classifications in six language versions of Wikipedia: Belarusian (be), German (de), English (en), Polish (pl), Russian (ru), Ukrainian (uk)

Grade / Language	be	de	en	pl	ru	uk
Featured Article (FA)	✓	✓	✓	✓	✓	✓
A-Class			✓			
Good Article (GA)	✓	✓	✓	✓	✓	✓
Solid					✓	
B-Class			✓			
Four				✓		
Full					✓	✓
C-Class			✓			
Developed					✓	✓
Start			✓	✓		
In development					✓	✓
Stub	✓		✓	✓	✓	✓

Wikipedia operates based on an open-editing model, meaning that anyone can create and edit the content of articles on the platform. As a result, information can be updated almost instantly, and the website can respond quickly to current events and discoveries. Community editing (including by anonymous users) allows errors to be corrected and for the content to be continually improved.

However, Wikipedia's open co-editing model also has its drawbacks. For example, encyclopedia articles are susceptible to deliberate misinformation and damaging content. Because it is not required for each and every change of content to be reviewed, harmful changes can immediately become visible to other Wikipedia readers. This inevitably leads to errors and inaccuracies in some texts. Moreover, Wikipedia articles can sometimes be biased, especially if edited by people with vested interests in a particular topic. In addition, the fact that anyone is able to edit articles can lead to conflicts between editors, over the specific content of articles or over how to interpret the rules.

Yet despite all these drawbacks, the overall philosophy of allowing anyone to edit Wikipedia has been, and indeed continues to be, key to its global success. Articles on this platform are co-created by volunteers from around the world, making it highly dynamic and able to keep up with the pace of events. More than half a million edits are made to Wikipedia every day, which means it would be very difficult to manually monitor all the changes.

Give that article a medal!

In each language, Wikipedia is created by a unique community of users, who shape and interpret the quality standards for their particular language version. Therefore, each version has slightly different criteria for content quality based on community discussion

DOI: 10.24425/academiaPAS.2023.147633

CLICKSDEMEXICO/SHUTTERSTOCK.COM

and experience. As a rule, each language version has special awards for articles that have achieved the highest quality standards.

In Wikipedia's most well-developed language version – English – the status of "featured article" (*FA*) is given to articles written in exemplary fashion, that meet all the quality criteria for this language version and so are worth emulating. "Good article" (*GA*), in turn, is a status given to articles that are close to meeting the standards of exemplary articles, but do not yet do so 100%. In the Polish-language version of Wikipedia, for instance, such content is referred to as *artykuł na medal* "top-notch article" (literally: "deserving of a medal") and *dobry artykuł* "good article" (two statuses analogous to the *FA* and *GA* classes in the English language version).

Before receiving one of these coveted distinctions, an article is subjected to careful and thorough scrutiny by the community. Users decide in open discussions whether a particular Wikipedia article meets the established criteria. Everyone can present arguments for and against awarding a particular status. Notably, the rules for awarding these distinctions can evolve and may be adapted to the needs of a specific language version. Such changes can lead to situations where certain articles may lose their previously granted status.

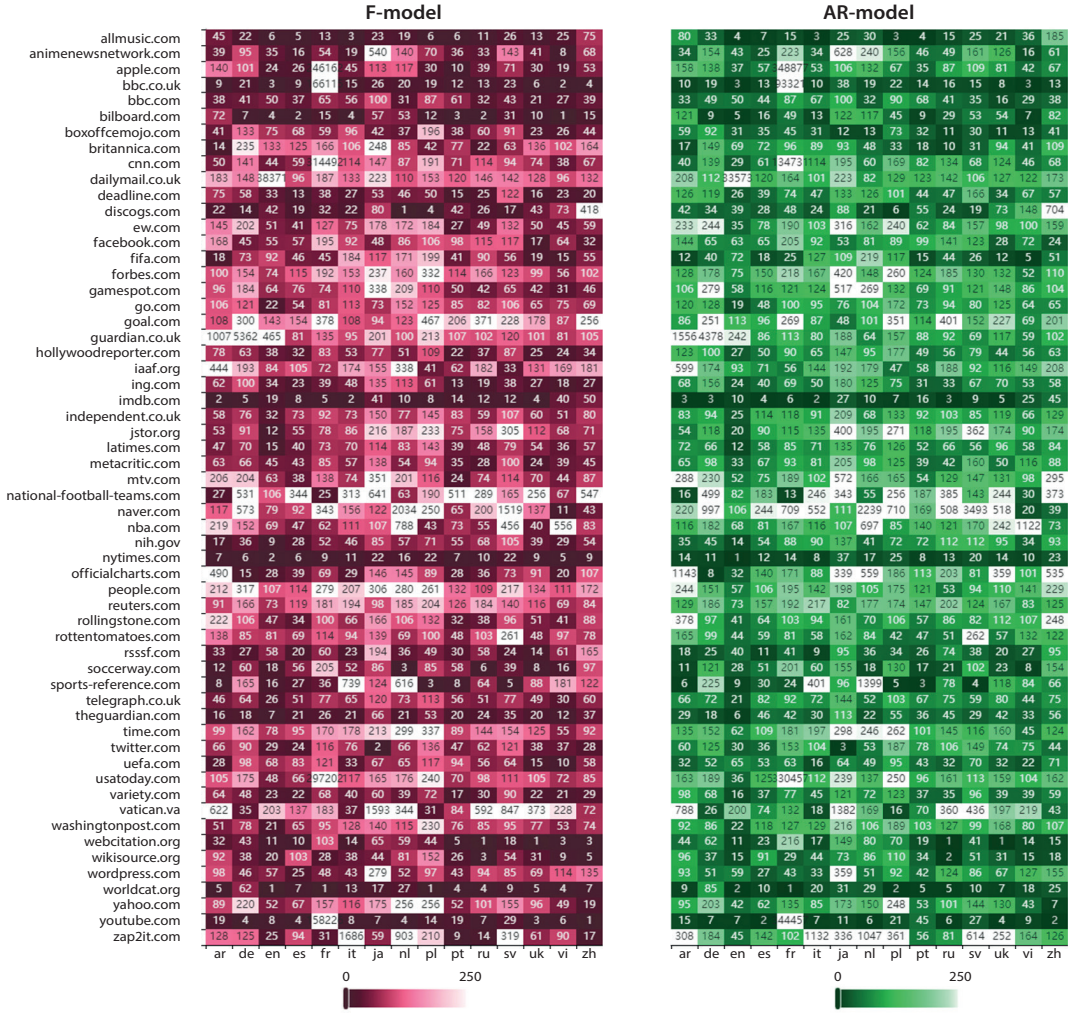
Some language versions of Wikipedia have a more developed quality-rating system that indicates how close an article is to achieving model status. In the English Wikipedia, articles are classified in seven quality categories, from highest to lowest: *FA*, *GA*, followed by *A-class*, *B-class*, *C-class*, *Start*, *Stub*. It is worth noting that quality grades lower than *FA* and *GA* can be individually assigned by users, without the need for discussion or community consensus. In the Polish Wikipedia, apart from the highest distinctions, articles are generally classified as: *czwórka* ("four," borrowed from the term for "B-grade" in school), *start*, and then *zalążek* ("stub"). Moreover, even within the same language version, different thematic sections of the encyclopedia may employ different names for similar quality grades.

The overall picture, therefore, is that while Wikipedia has standards for quality assessment, these criteria can vary depending on the language version and also may change over time. Moreover, judgements are often subjective, requiring collaboration and agreement among editors. All this means that automating the content-quality assessment process in Wikipedia could greatly contribute to improving the credibility and efficiency of edits. Algorithms, after all, operate often on quantitative measures leading to more consistent assessments in more objective way, without emotional or subjective interference. This will also allow for the automatic collection of large amounts of data to get quality measures for billions of documents and faster identification of problems related to their content.

Computer tools can quickly identify vandalism or misinformation, providing editors with up-to-date information and suggesting corrections. Additionally, in the case of deliberate disinformation attacks, such tools can act as a defensive mechanism.

Our team at the Department of Information Systems at the Poznań University of Economics and Business is engaged in research on automated evaluation of the quality of Wikipedia articles. Some of the models we have developed and reported in scientific publications have been implemented on publicly accessible websites. For example, the WikiRank.net project allows users to check the quality and popularity of Wikipedia articles in different language

MN STUDIOS/SHUTTERSTOCK.COM



The aggregate rankings of the most important sources of information cited by Wikipedia articles related to culture in the broadest sense of the term

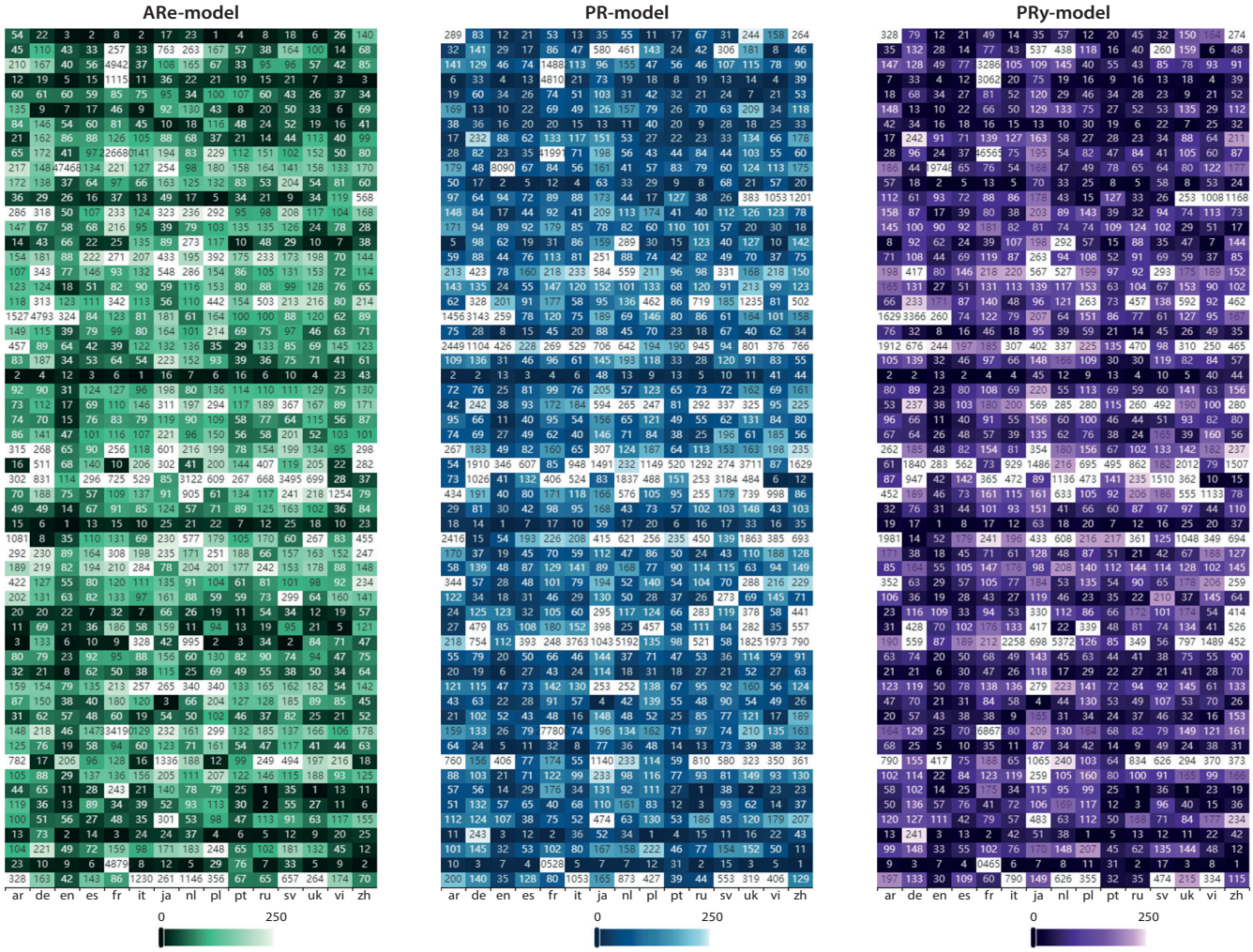
versions. Another example is the BestRef.net project, for analyzing the importance of information sources on Wikipedia.

Reliable information sources

A key aspect of Wikipedia article quality is the principle of verifiability of information. This means that the information in the encyclopedia articles must be based on reliable, credible sources. However, how source reliability is judged can vary, depending on the specific subject of the article and the Wikipedia language version. Additionally, the reliability of a source depends on factors such as the reputation of the publisher or author, the review process, and the accuracy of the information presented. When evaluating sources to be cited in Wikipedia, editors should aim to choose those with a good reputation and are widely recognized as reliable in their fields. However, the biggest challenge posed by the concept of source reliability (much like the concept of information quality) is the subjectivity of the evaluation process. This means that Wikipedia editors must reach a consensus on every information source that can be used in Wikipedia articles.

Only a few of the well-developed language versions of Wikipedia provide even a non-exhaustive list of sources whose reliability and application in Wikipedia have often been debated. Even the English Wikipedia (the largest version) has such a general compilation of reliability information for only about 400 websites. Sometimes we can find similar lists for specific topics (e.g., video games, movies, new articles in English Wikipedia). Given that there are over a billion websites on the web, assessing the reliability of each of them individually would pose a huge challenge. Additionally, it must be borne in mind that the reputation of the very same source can change over time, which may require further regular reliability assessment.

Automating the source-assessment process can help quickly identify sources that are potentially unreliable, outdated, or do not meet academic standards, allowing editors to focus on scrutinizing them or replacing them with more credible sources. Additionally, in this age of rampant misinformation, auto-



matic source-assessment can quickly detect and flag information that is based on dubious sources, preventing it from spreading. Moreover, new Wikipedia editors may not be sure which sources are most reliable in a given field. Automatic source assessment can provide them with guidelines and recommendations, helping them to choose appropriate source materials.

A study of all Wikipedia articles in all the different language versions showed that there are over a million different websites that are used in over 300 million references of Wikipedia articles. In the most developed language version (English), nearly 77 million references can be identified, and about 8 million in the Polish-language version. Using various models for assessing Internet sources, we can identify the most important ones from the standpoint of individual language versions of Wikipedia.

Semantic databases

The advancement in semantic technologies has greatly enhanced the efficiency of processes like information retrieval, sentiment analysis, and content summarization. Two good examples here are the platforms DBpedia and Wikidata. DBpedia transforms Wikipedia data into a format more accessible for machines, while Wikidata acts as a centralized database for all Wikimedia projects in various languages. These platforms not only facilitate structured knowledge access but can also be used to improve the overall quality of Wikipedia in different languages. On the other hand, higher quality content on Wikipedia also contributes to higher quality in these semantic knowledge bases.

Wikipedia, Wikidata, and DBpedia are open resources that allow their content to be used for various purposes. Better quality of these resources, in turn, can contribute to improving other a wide array of services and applications that use open data, including: internet search engines, natural language processing applications, educational applications, recommendation systems, virtual assistants, cultural and tourism applications, network connections, and many more. ■

Further reading:

Lewoniewski W., Węcel K., Abramowicz W., Companies in Multilingual Wikipedia: Articles Quality and Important Sources of Information, *Lecture Notes in Business Information Processing*, 2023.

Lewoniewski W., Identification of Important Web Sources of Information on Wikipedia across various Topics and Languages, *Procedia Computer Science*, 2022.

Lewoniewski W., Węcel K., Abramowicz W., *Reliability in Time: Evaluating the Web Sources of Information on COVID-19 in Wikipedia across Various Language Editions from the Beginning of the Pandemic*, 2022.

Lewoniewski W., PhD thesis available at: https://www.wbc.poznan.pl/Content/461699/Lewoniewski_Wlodzimierz-rozprawa_doktorska.pdf