





Wiki Evolution dataset applicability: English Wikipedia revision articles represented by quality attributes

Ana Luiza Sanches  [Centro Federal de Educação Tecnológica de Minas Gerais | analuiatrz@gmail.com]

Sinval de Deus Vieira Júnior   [Centro Federal de Educação Tecnológica de Minas Gerais | sinvalvieirajunior@gmail.com]

Daniel Hasan Dalip  [Centro Federal de Educação Tecnológica de Minas Gerais | hasan@cefetmg.br]

Bárbara Gabrielle C. O. Lopes  [Universidade Federal de Minas Gerais | barbaragcol@dcc.ufmg.br]

 Departamento de computação, Centro Federal de Educação Tecnológica de Minas Gerais, Av. Amazonas, 7675 - Nova Gameleira, Belo Horizonte, MG, 30510-000, Brazil.

Received: 23 July 2023 • Published: 5 April 2024

Abstract This paper presents the creation of the Wikipedia article's evolution dataset. This dataset is a set of revisions of articles, represented by quality attributes and quality classification. This dataset can be used for studies regarding automatic quality classification that consider the article revision history as well as understanding how the content and quality of articles evolve over time in this collaborative platform. To illustrate a potential application, this study provides a practical example of utilizing a Machine Learning model trained on the constructed dataset.

Keywords: Wikipedia, Dataset, Information Quality

1 Introduction

The Internet allows a wide sharing of content by any user. Also, the increase in Internet users has made possible an enormous and growing volume of information hosted by content platforms. An example of a web content platform is Wikipedia, which has more than 5 million articles in English written through a collaborative effort involving 37.6 million registered users and an indefinite number of anonymous users [Wikipedia, 2023c].

Considered one of the largest repositories of human knowledge, Wikipedia received a lot of attention and the quality assessment of its articles became a major concern during the 2000s [Dang and Ignat, 2016]. The article quality concern is mainly due to the discussion of collaborative texts reliability, because, as a consequence of its open structure, Wikipedia cannot guarantee, in any way, the validity of the information contained therein [Wikipedia, 2023b]. Therefore, it is considered essential to refer to external sources and “inline” citations to verify the information contained in the articles. Currently, Wikipedia article classification involves volunteer reviewers responsible for classifying articles into seven quality classes [Wikipedia, 2023a]. The quality classes are, in descending order, *Feature Article* (FA), *Good Article* (GA), A, B, C, *Start* and *Stub*.

To identify the article quality, however, human experts are not enough, because the high speed of change in articles makes it impossible to perform this task manually [Dang and Ignat, 2016]. Because of that, several works have explored the automatic classification of articles taking into account several criteria such as collaborative contribution, identification of vandalism, identification of controversy, user feedback, among others [Jhandir *et al.*, 2017].

In all these works, article data collection and structuring is

an essential step. Collected data can contain various issues such as missing values, false data, duplicate data and lack of standardization [Batista *et al.*, 2018]. Such problems are usually solved in the data pre-processing step, which requires about 80% of data scientists' time [Tyagi *et al.*, 2010].

Thus, the present work aimed to create and provide a database that can be used to obtain automatic methods of predicting the quality of Wikipedia articles. The collection algorithm used to create this database is also being made available. The main contribution of this dataset is the extraction of text from the Wikipedia revision history. We hypothesize that the evolution of the article may be useful to assess the quality of the article. Therefore, this dataset contains a representation of the evolution of a review, which uses quality indicators of the article to be classified and a previous version of it. We have extracted two samples: a smaller dataset with 3,246 and approximately 53,000 reviews (the articles used in Hasan Dalip *et al.* [2009]) and a bigger one containing 35,572 articles and 2,175,236 reviews.

Then, to assess the effectiveness of the dataset, a practical experiment was conducted, wherein a Support Vector Machine (SVM) model was applied to predict the evolution of articles in Wikipedia. The goal is to demonstrate the practical application and efficacy of the dataset in predicting article quality. Additionally, the experiment provided insights into the challenges encountered when implementing the dataset in a Machine Learning context¹.

This article is structured as follows: Section 2 presents the related work. Section 3 presents how Wikipedia makes its data available and how it was used to generate the dataset,

¹Part of this work have been published, in Portuguese, in DSW'2022: Sanches, A. L.; Júnior, S. D. D. V.; Dalip, D. H.; Lopes, B. G. C.. Wiki Evolution dataset: English Wikipedia revision articles represented by quality attributes. In Proceedings of the IV Dataset Showcase Workshop (DSW).

Section 4 describes the collection algorithm developed, specifying parameters and results of each step. Then, Section 5 presents the collection of two sample datasets which are available for use. Finally, the section 6 presents a practical experiment with the dataset and some possible other uses, besides future work improvements.

2 Related Work

To emphasize the importance of this dataset, and which features are important to extract, in this section we present several works regarding the automatic classification of Wikipedia articles.

Blumenstock [2008] represents Wikipedia articles using a very simple approach which only considers article size to rank their quality. Therefore, the binary model is based on predicting the article’s quality class, whether it has a FA class, and is based on size measured in words. Lipka and Stein [2010] also proposes a binary classification, classifying FA or non-FA, but represents articles by their writing style. For that, attributes extracted through the Bag-Of-Words and trigram techniques were used.

Other articles define notions of quality more specifically, such as Dondio *et al.* [2006], which proposes an automatic and transparent mechanism that estimates the reliability of Wikipedia articles. To this end, the author analyzes the Wikipedia domain, creating reliability features compatible with the collaborative nature of the repository. This study has divided the features into two domains: content quality (CQ) and collaborative editing (CE). This work uses mathematical formulas to assess the article’s quality.

Dalip *et al.* [2011] makes use of a regression model based on *Support Vector Machines* (SVM) to solve the problem of automatic article qualification. This model, known as *Support Vector Regression* (SVR) outputs a single value on a continuous quality scale, unlike the Dondio *et al.* [2006] model which is a binary classifier. In this work, the attributes are indicators represented by statistical measures correlated with quality. A total of 54 indicators were created, and distributed among the groups review attributes (13), network attributes (8), text attributes (18), style attributes (9), and readability attributes (6).

Warncke-Wang *et al.* [2013] also uses quality features as indicators and analyses them to identify those that contribute most to quality prediction. Among the attributes studied, those that stood out the most were article size, number of references/article size, number of sections, completeness (measurement of the number of links to another article), and informativeness (number of images + noise).

Several researchers create prediction models to determine the quality class for an article without taking the previous class into account, such as Raman *et al.* [2020a], Ruprechter *et al.* [2020], Raman *et al.* [2020b], Ma *et al.* [2017]. Specifically, we will talk deeply about Wang and Li [2020] and Sugandhika and Ahangama [2022].

Wang and Li [2020] utilized the same indicators as Dalip *et al.* [2011] to represent articles at a specific time, aligning with our research. However, they did not consider the article evolution, having just one version of each article. In

contrast, our research acknowledges that an article’s history may have a few number of class transitions and we substantiate the impact of this observation through the presentation of our results.

In the work by Sugandhika and Ahangama [2022], a Wikipedia dataset for content assessment was generated, and an EAT model was employed to construct a prediction model. Notably, the study did not incorporate an analysis of the articles’ evolutionary changes in the development of their model, a facet which we address and emphasize in our current study.

Dang and Ignat [2016] addresses the automatic quality assessment of Wikipedia articles, dividing the studies into two groups: those that use revision history and those that use article content. Revision history studies consider attributes such as which people edited the revisions, contributed and analyzed edits in other revisions. Article content studies contain features considering its length, complexity, number of images, presence of information box, and organization into sections, not making use of metadata. Examples of these groups are works by Blumenstock [2008] and Lipka and Stein [2010]. There is also a third group, which consists of the union between these two families, that is, attributes based on the history and content of the article, an example being the work of Dalip *et al.* [2011]. However, as far as we know, our study is the first which uses the evolution of the article and, consequently, their previous class, in order to predict the article.

3 Wikipedia organization and data access

In order to perform the crawling process, first it is important to understand the organization of Wikipedia and how the data is made available for consumption. Thus, this section presents the organization of Wikipedia and the alternatives to collect data from it.

Wikipedia articles have four web pages which are important to highlight in this work, as they present the data necessary for creating the dataset: the article page, talk page, article page revision history, and talk page revision history.

Each article has a talk page (Figure 1) through which editors and reviewers discuss changes made to the article content. This page also contains the article quality class.

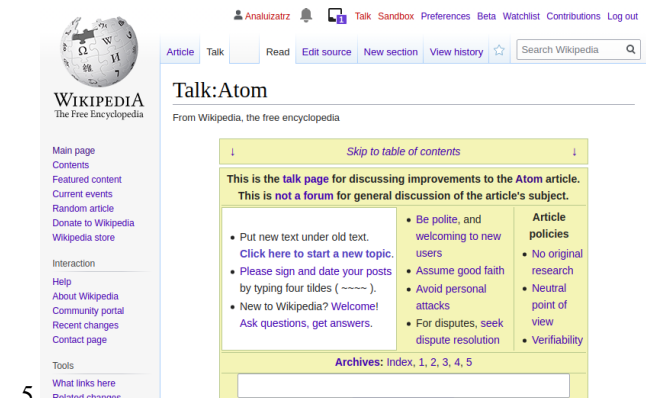


Figure 1. Atom talk page

Article revision pages display past versions of a particular

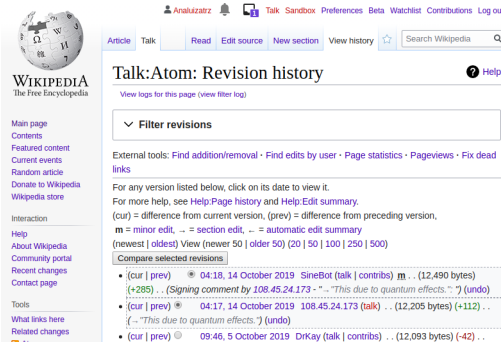


Figure 2. History revision page for Atom talk page

article. From them, it is possible to access the content of an article on a given date. In this work, revision is defined as a version of the article on a date. Therefore, the article page displays the most current revision of the article, while the revision page displays the history of previous versions.

Similarly to the article page, talk pages also have a revision history, making it possible to query past discussions regarding article revisions. The talk page presents, among other data, the quality class, making it possible to obtain the quality of previous revisions of an article through the article’s revision history page (Figure 2).



Figure 3. Atom article quality evolution
Source: Wikipedia [2023a]

Talk pages allow us to access past versions of content and allow us to extract the quality evolution of articles over time. Figure 3 shows the evolution of the “Atom” article over time, using Wikipedia’s quality classes.

According to the Wikipedia, these are the following grades²

- **Featured Article (FA):** According to their evaluators, the best Wikipedia articles. It can be described as professional, outstanding, and thorough, a definitive source for encyclopedic information.
- **A-Class:** these are articles considered to be well organized and essentially complete, being very useful to readers. But with a few pending issues that need to be solved.
- **Good Article (GA):** The article meets all of the good article criteria³: Well-written, Verifiable, Broad in its coverage, Neutral, Stable, Illustrated, if possible. It is useful to nearly all readers, with no obvious problems; approaching (though not necessarily equalling) the quality of a professional publication.
- **B-Class:** Mostly complete articles with no major problems, but requires some further work to reach good article standards. Readers are not left wanting, although the content may not be complete enough to satisfy a serious student or researcher.
- **C-Class:** The article is substantial but is still missing important content or contains irrelevant material. The

article should have some references to reliable sources, but may still have significant problems or require substantial cleanup. Useful to a casual reader, but would not provide a complete picture for even a moderately detailed study.

- **Start-Class:** An article that is developing but still quite incomplete, citing or not citing adequate reliable sources. Provides some meaningful content, but most readers will need more.
- **Stub:** An elementary description of the topic. that meets none of the Start-Class criteria. Provides very little meaningful content and may be little more than a dictionary definition. Readers probably see insufficiently developed features of the topic and may not see how the features of the topic are significant.

Note that in Wikipedia, there is a quality classification for lists of articles (i.e. Feature List). However, this work did not take into account Wikipedia article lists, as they are a different type of document in which the quality criteria and, consequently, the features would not be the same as Wikipedia articles.

3.1 Data consumption and representation

There are two main options available for consuming data from Wikipedia, via the API⁴ and by collecting raw HTML pages. We opted for consuming data via APIs, as these tools have the advantage of a well-defined format and a clear schema that are generally well-documented so that their users can use them [Batista et al., 2018].

Even having a well-defined API, many systems do not specifically meet user demand. This means that APIs can make data available in different formats than what is needed for use, being necessary to implement a specific application to consume the data and structure it in the desired format [Batista et al., 2018].

From the MediaWiki API, for example, it is possible to consume the revision history of the article and the talk page. For this work, the content of each article revision is used to generate quality attributes related to size, style, structure, and readability [Hasan Dalip et al., 2009]. The final dataset contains 44 quality attributes, described in Dalip [2015]. The revision history of the talk page is used to extract the quality class. After that, it is necessary to join the quality features to their respective quality class.

4 Crawling algorithm

The crawling algorithm is a pipeline which aims to create the evolution database. The algorithm takes as a parameter a list of titles and a period of time and returns as a result an evolution base in the form of a CSV file. The period is determined by a start date and an end date. The algorithm is available at <https://github.com/analuzatrz/wiki-crawler>.

Figure 4 presents the pipeline steps, which use the Wikimedia API through HTTP requests, and the data handling

²https://en.wikipedia.org/wiki/Wikipedia:Content_assessment#Grades:

³https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria#The_six_good_article_criteria

⁴APIs, which stands for “Application Programming Interface” are well-defined interfaces for consuming data on the Web

steps, which extract or transform the response from the requests. The rectangles represent crawling, while the circles represent data processing.

The first crawler goal was to extract the review metadata presented in Table 1. Once collected, the metadata was filtered to select only one instance per month for the defined period. The selected metadata is the latest of each month. This approach allows us to reduce the collected data without losing so much relevant data, as an article’s quality class doesn’t change frequently.

Table 1. Revision dataset instance metadata

Metadata	Description
ID	revision identification
parent ID	parent revision identification, prior revision which does not contain the actual changes
comment	revision changes description
timestamp	revision creation date
access	fictitious date of access to the article
user	revision author

From the metadata selection, two other stages of the pipeline are carried out, the first to obtain the article page and the second to obtain the talk page. Quality attributes are extracted from article pages and quality classes are extracted from talk pages. Each metadata is the representation of a review, mainly characterized by title and date, which helps in future crawling. After the crawling and extraction steps, each review is characterized, in addition to the title and date, by a vector of quality attributes of 44 dimensions and a quality class.

Algorithm 1 shows an overview of how to extract the revision dataset. From the collected metadata, the attributes, and the class are obtained, which make up the review base instance. The development of the algorithms for each step of the collection, collectMetadata(), collectArticlePage(), and collectTalkPage() was iterative. This means that a sample collection was performed while the algorithms were being improved. This sample database will be in the presented section 6.

Algorithm 1: Article collection that results in the revision dataset

Data: titles, startDate, endDate
Result: R
 $R \leftarrow \emptyset$;
 $M \leftarrow \text{collectMetadata}(\text{titles}, \text{startDate}, \text{endDate})$;
foreach $m_t^p \in M$ **do**
 $W_p \leftarrow \text{collectArticlePage}(m_t^p)$;
 $\mathbf{a}_t^p \leftarrow \text{extractAttributes}(W_p)$;
 $W_d \leftarrow \text{collectTalkPage}(m_t^p)$;
 $c_t^p \leftarrow \text{extractClass}(W_d)$;
 $r_t^p \leftarrow [\mathbf{m}_t^p, \mathbf{a}_t^p, c_t^p]$;
 $R \leftarrow R \cup \{r_t^p\}$;

The algorithm has evolved to incorporate error recovery mechanisms as well as handling disambiguation and redirects. For error recovery, a record was created of the pages already collected and the pages that were not collected due to an error. When any collection step starts, articles already collected are not placed in the collection queue again. It is also possible to study the articles that were not collected and understand why the error occurred. Redirection happens when

the page accessed is a redirect page⁵, which automatically sends the visitor to another page. Redirect pages are useful for referencing the same article by another title (e.g. Einstein⁶ and Albert Einstein⁷ are redirected to the same content. Disambiguation, in turn, happens when a title is ambiguous and can refer to two or more distinct articles.

The disambiguation pages⁸ are useful to help the visitors find the article that interests them by displaying the articles related to the ambiguous title (e.g. there are three articles linked to the title "Mercury"⁹, the element¹⁰, the planet¹¹ and mythology¹²).

4.1 Metadata Crawling

The metadata crawling algorithm receives as parameters the article titles and the period of time. The period of time is represented by initial and final dates in which the revisions were made. The information obtained by this crawler is revision id, page id, revision date, user who performed the modification, and revision comment. The metadata is saved in a CSV file to be later used by the next crawlers.

4.2 Article Page Crawling Process

The article page crawling process consists of fetching content from the article pages of each review. The algorithm has as parameters the pair title and revision date, from its metadata.

After collecting the content of the article pages of each review the attributes are extracted. Before the extraction, the article page content must be converted. The original format of the pages is a Wikipedia format called Wikitext¹³ and the target is a HTML page. From the HTML version the quality attributes with the web quality library [Pinto *et al.*, 2020].

4.3 Crawling the Talk Page

Talk page revisions and article page revisions are not necessarily synced, which means that they might be edited at different times. In order to match talk page revisions with article page revisions, it was considered for a given article page version their later closest date of the talk page. By doing this, we ensure that the talk page revision associated with an article revision is the one that has the earliest possible date considering the article version.

The talk page content crawler has as parameters the pair title and revision date. Article page revisions are not directly

⁵<https://en.wikipedia.org/wiki/Wikipedia:Redirect>—accessed on November 15th, 2019.

⁶<https://en.wikipedia.org/wiki/Einstein>. Accessed in November 15th, 2019.

⁷https://en.wikipedia.org/wiki/Albert_Einstein. Accessed on November 15th, 2019.

⁸<https://en.wikipedia.org/wiki/Wikipedia:Disambiguation>. Accessed in November 15th, 2019.

⁹<https://en.wikipedia.org/wiki/Mercury>. Accessed in november 15th, 2019.

¹⁰[https://en.wikipedia.org/wiki/Mercury_\(element\)](https://en.wikipedia.org/wiki/Mercury_(element)). Accessed in november 15th, 2019.

¹¹[https://en.wikipedia.org/wiki/Mercury_\(planet\)](https://en.wikipedia.org/wiki/Mercury_(planet)). Accessed in november 15th, 2019.

¹²[https://en.wikipedia.org/wiki/Mercury_\(mythology\)](https://en.wikipedia.org/wiki/Mercury_(mythology)). Accessed in november 15th, 2019.

¹³<https://en.wikipedia.org/wiki/Help:Wikitext>

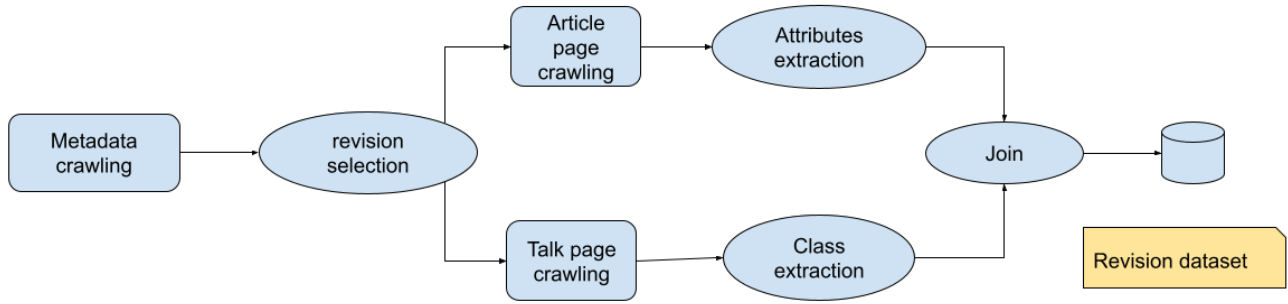


Figure 4. Diagram of data collection and treatment to create the review database

related to talk page revisions. In this work, it was considered that since the quality is evaluated on an already written revision, the discussion page considered is the one whose date is the closest posterior to the revision date. For example, given an article page review of April 5th and two talk page reviews on April 4th and 6th, quality considered will be drawn from the April 6th review. These parameters are obtained from metadata discretized month by month.

From the Talk Page content, the revision quality class is extracted. Because wiki pages have their markup language format known as Wikitext, the class detection was done using a regular expression (regex) capable of identifying the following pattern `"{{project|class=x|prop_1=?|(...)}}`"¹⁴, being x the class which we want to extract. The regex detects to extract the class is the following `"class()*=()*([a-z]+)"`. Figure 5 shows an example of a Talk Page for *Binary Search* article.

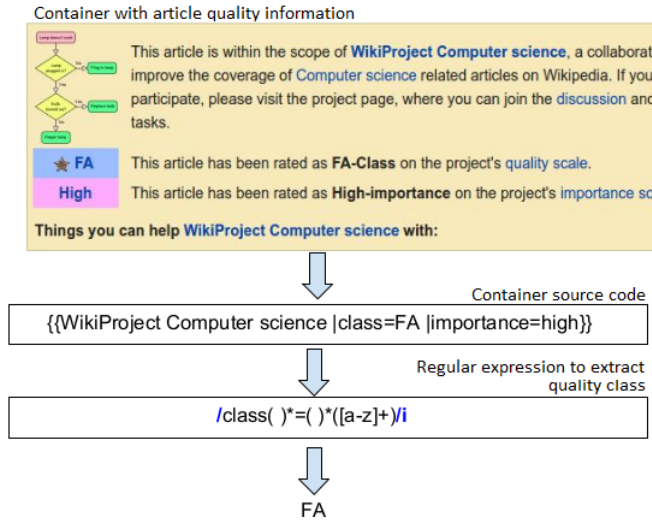


Figure 5. Quality class extraction demonstration, given an article Talk Page (*Binary Search*).

Source: Batista et al. [2018]

After quality attributes and class extraction from each revision, the vectors metadata \mathbf{m}_t^p , attributes \mathbf{a}_t^p and class c_t^p , from each revision, are concatenated to build a dataset instance \mathbf{r}_t^p . As the revision can be identified by a page (article title) p and date (timestamp) t , a dataset revision instance R

¹⁴https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Avalia%C3%A7%C3%A3o_autom%C3%A1tica

Table 2. Initial sample statistics

Collection	# Articles	Errors	Total
Metadata	3,246	48	3,294
Article Page	3,242	4	3,246
Quality class	2,999	247	3,246

can be represented by the Equation 1.

$$\mathbf{r}_t^p = [\mathbf{m}_t^p, \mathbf{a}_t^p, c_t^p] \forall p \in P, \forall t \in T \quad (1)$$

5 Dataset

By using the crawling process described in Section 3 we created two samples. The first goal is to use the same dataset instances as in Hasan Dalip et al. [2009] to allow easier comparisons to previous works. The second sample was created using a larger number of titles during a longer period, aiming to create a more comprehensive database. The initial sample was possible to record the execution time and storage of intermediate documents. The final sample was collected fractionally in an effort that lasted months and we did not record the exact execution time of each step.

5.1 Initial sample collection

To develop the initial sample database, we used as parameters the 3,294 articles from 2007 to 2009 presented in Hasan Dalip et al. [2009]. Table 2 shows details of the collection stages and in how many titles errors occurred. These errors concern titles which were not available anymore as it was renamed or deleted.

In the collection of article pages, 53,023 reviews of 3,242 pages were obtained, consuming 1.5 GB of storage. To collect these data, the execution time of the algorithm was approximately 5 hours and 17 minutes.

After collecting the content of the article pages of each review, attribute extraction was performed. One of the attribute extraction steps is converting the wiki format to HTML, which resulted in the persistence of approximately 53,000 pages consuming 2.2 GB of storage. The execution time for this conversion was approximately 3 hours.

In total, we collected discussion pages from reviews of 2,998 articles, which make up a database of 58,024 instances of 70 attributes.

5.2 Final sample collection

First, more than 6,000,000 article titles were collected. Table 3 presents, among other information, the distribution of titles by quality class, as of March 2020. As shown in this table, some classes have a very small sample, such as class A with less than 0.02% of titles.

Table 3. Articles by Quality Class

Class	Total of Articles (titles)	Sample Articles	Sample reviews
FA	5,750	5,750	609,502
A	1,072	1,072	7,053
GA	11,689	5,750	348,257
B	100,410	5,750	450,237
C	328,210	5,750	359,277
Start	1,691,461	5,750	225,451
Stub	4,085,974	5,750	111,979
Total	6,224,521	35,572	2,175,236

To make the final database more balanced, some titles were disregarded, in a subsampling process. For this, 5,750 articles of each class were randomly selected from the collected titles. We limited to 5,750 as this is the total of Wikipedia Feature Articles. Also, we opted to use all the Feature Articles since it is the final quality class and, because of that, these articles contains a more representative overview of the evolution of an article. As in A-Class there were only 1,072 titles available, we collected all of them.

We executed the algorithm using a sample resulting in a database of more than 2,000,000 reviews. The chosen reviews were from January 2003 to March 2020, which configures the entire period of data available when collection began. Table 3 (Sample Articles column) presents the distribution of these reviews by class. The classes with the highest sample are B, FA and Start. The generated database is publicly available at https://figshare.com/articles/dataset/FinalWikiEvolutionSample_csv/20154434 and the column description is at <https://github.com/analuzatrz/wiki-crawler/tree/master/FinalWikiEvolutionSample>.

5.3 Limitations

The dataset created has some limitations. Regarding the month selection, some reviews were grouped. This is due to the choice to prioritize a dataset with revisions over a long period rather than all revisions over a shorter period, which could affect some applications of the dataset. Another limitation is the representation of quality based only on the content of the review. A possible improvement in this sense would be to enrich the database with other attributes, such as article authorship, which can be crawled based on authorship metadata of reviews and incorporated into the representation of articles.

Futhermore, the Wikipedia manual quality classes assessment may have bias. To minimize this, the Wikipedia community organizes the article in WikiProjects which is a collection of articles that belongs to the same category (e.g. history, geography, insects, etc). Then, the community tries to assess the quality by considering the specificities of their WikiProject. Furthermore, FA and GA pass through a more rigid qual-

ity assessment^{15,16}. Then, they are assigned as a FA or GA just when the community has reached a consensus¹⁷ that the article meets the quality class criteria. Another limitation is that this work did not consider Wikipedia lists, as it would be a different kind of assessment with different criteria and features.

6 Applications

The created dataset can serve as a database for several studies. Among them, studies that seek to analyze the evolution of human writing, especially during the 21st century, will be able to use the proposed methodology to generate a comprehensive sample that represents the evolution of Internet content. Thus, it will be possible to analyze which aspects of writing were more developed, in addition to allowing the study of which of these aspects most influence the overall quality of an article, and, in this way, create tools that automatically determine its quality and make relevant suggestions for improvements.

With this in mind, this work also includes a practical experiment that utilizes the dataset created using the methodology presented in this study. The experiment aims to demonstrate the application and effectiveness of the dataset in predicting the evolution of articles in Wikipedia. Moreover, it sheds light on the challenges encountered when applying the dataset in a Machine Learning context, highlighting the complexities and biases that arise in the prediction process. By addressing these challenges, this experiment contributes to the broader understanding of utilizing such datasets and paves the way for future research in this domain.

6.1 Quality Prediction Experiment

To show the feasibility of predicting the quality of an article by considering their past review, we applied a Linear Support Vector Machine (Linear SVM) model to predict the quality evolution of articles on Wikipedia in our dataset. To accomplish this, we have used 300,000 revisions from our dataset due to performance issues. Out of these, 100,000 revisions were used to determine the cost parameter (vary within the range of 1, 10, 100 and 1000), while the remaining 200,000 revisions were used for training the model. We divided the dataset in train, test and validation. To evaluate the results, we computed the accuracy scores.

In this study our goal is **G1**: to present the impact of using the previous classes to predict the current quality classes; and, also, **G2**: to investigate the challenges when using the previous class

For the first goal, we analyzed the performance of the SVM model considering the class from the previous revision and comparing when not using it. By doing this, we can understand how the evolution contributes to the quality prediction accuracy. Then, Table 4 presents the prediction accuracy comparison when using or not the previous class. Since the previous class is an important feature, their usage could

¹⁵<https://en.wikipedia.org/wiki/WP:FAC>

¹⁶<https://en.wikipedia.org/wiki/WP:GAC>

¹⁷<https://en.wikipedia.org/wiki/Wikipedia:Consensus>

greatly improve performance. However, a deeper investigation is needed to understand why we could reach this result (G2).

Table 4. Accuracy impact when using the previous class

	Considering previous class	Not considering prev. class
accuracy	0.85	0.465

Thus, first we analyze in Figure 6 which presents the number of reviews according to the previous class (column position) and the current class (row position). Therefore, the principal diagonal represents the class permanence, and the rest are class transitions. By doing this, we can observe that the class remains the same in most of the cases. This can be considered a natural behavior since an article takes many revisions to obtain a new quality class. However, this poses a challenge for machine learning prediction models, as this aspect introduces a bias that may lead the model to assume the next class of an article is the same as the previous one.

		Current class						
		FA	GA	A	B	C	Start	Stub
Previous class	FA	348,979	11	25	65	14	5	5
	GA	2,846	220,747	1,113	152	48	9	1
	A	852	152	28,656	120	9	2	0
	B	790	4,060	288	345,842	815	531	29
	C	138	1,293	35	1,347	182,262	178	10
	Start	247	1,589	71	2,802	2,777	335,341	359
	Stub	55	327	18	1,058	997	3,173	177,094

Figure 6. Number of reviews according to the previous class (row position) and the current class (column position)

To understand the impact of fewer class transitions in the model, we examined the model's performance specifically for transitions between classes. To accomplish this, Table 5 presents the results of the dataset grouped by the reviews with and without the transitions.

When there is no transition between classes, the model achieved an accuracy score of 99.59% when considering the previous class, and 43.76% when not considering it. These results highlight the model's effectiveness in predicting the evolution of articles when no class transitions occur. However, it's worth noting that when analyzing the dataset with transitions, the accuracy score dropped significantly to 5% when considering the previous class, and 45.76% when not considering it. These findings confirm a clear bias in the model's performance, particularly when providing the previous class information.

Table 5. SVM - Results by transitions

	Without transitions	With transitions
Considering previous class	0.9959	0.05
Not considering prev. class	0.4376	0.4576

7 Conclusion

This article presents a Wikipedia Dataset that takes into consideration the evolution of the article. We also presented the feasibility of predicting the quality of an article by considering past reviews of the article. Nevertheless, it is worth

emphasizing that this experiment holds significance in guiding future research endeavors and the practical application of the dataset developed in this study. By recognizing the limitations of the SVM model and emphasizing the need for further exploration, this research contributes to the broader understanding of predicting article evolution and lays the foundation for more advanced methodologies in the future.

Then, in future studies, it is crucial to explore sequence models, such as Long Short-Term Memory (LSTM) and attention models, to improve results. The adoption of a sequence model can help address the observed bias in the SVM model and facilitate more accurate predictions.

References

- Batista, N. A., Brandão, M. A., Pinheiro, M. B., Dalip, D. H., and Moro, M. M. (2018). Dados de múltiplas fontes da web: coleta, integração e pré-processamento. In de Computação – SBC, S. B., editor, *Anais do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web: Minicursos*, chapter 5, pages 153–192. Sociedade Brasileira de Computação – SBC.
- Blumenstock, J. E. (2008). Size matters: Word count as a measure of quality on wikipedia. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 1095–1096, New York, NY, USA. ACM. DOI: 10.1145/1367497.1367673.
- Dalip, D. H. (2015). *Uma Abordagem Multi-Visão para a Estimativa Automática da Qualidade de Conteúdo Colaborativo na Web 2.0*. PhD thesis, UFMG.
- Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2011). Automatic assessment of document quality in web collaborative digital libraries. volume 2(3), page 1–30. DOI: 10.1145/2063504.2063507.
- Dang, Q. V. and Ignat, C.-L. (2016). Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL '16*, pages 27–30, New York, NY, USA. ACM. DOI: 10.1145/2910896.2910917.
- Dondio, P., Barrett, S., Weber, S., and Seigneur, J.-M. (2006). Extracting trust from domain analysis: A case study on the wikipedia project. volume 4158, pages 362–. DOI: 10.1007/11839569_35.
- Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '09*, pages 295–304, New York, NY, USA. ACM. DOI: 10.1145/1555400.1555449.
- Jhandir, M. Z., Tenvir, A., On, B.-W., Lee, I., and Choi, G. S. (2017). Controversy detection in wikipedia using semantic dissimilarity. *Inf. Sci.*, 418(C):581–600. DOI: 10.1016/j.ins.2017.08.037.
- Lipka, N. and Stein, B. (2010). Identifying featured articles in wikipedia: writing style matters. pages 1147–1148. DOI: 10.1145/1772690.1772847.
- Ma, Z., Tao, J., and Hu, J. (2017). The dynamics of wikipedia

- article revisions: an analysis of revision activities and patterns. *International Journal of Data Mining, Modelling and Management*, 9(4):298–314.
- Pinto, A. C., Silva, B. S., Carmo, P. R. M., Lima, R. L. A., Amorim, L. S. P., Viana, R. T. C., Dalip, D. H., and Oliveira, P. A. C. (2020). Webfeatures: A web tool to extract features from collaborative content. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 103–106, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/webmedia_estendido.2020.13071.
- Raman, N. et al. (2020a). Classifying wikipedia article quality with revision history networks. In *Proceedings of the 16th International Symposium on Open Collaboration*.
- Raman, N. A. R. U. N. et al. (2020b). Revisionist history: Predicting wikipedia article quality with edit histories.
- Rupprechter, T., Santos, T., and Helic, D. (2020). Relating wikipedia article quality to edit behavior and link structure. *Applied Network Science*, 5:1–20.
- Sugandhika, C. and Ahangama, S. (2022). Assessing information quality of wikipedia articles through google’s e-a-t model. *IEEE Access*, 10:1–1. DOI: 10.1109/ACCESS.2022.3172962.
- Tyagi, N., Solanki, A., and Tyagi, S. (2010). An algorithmic approach to data preprocessing in web usage mining. *International Journal of Information Technology and Knowledge Management*, 2.
- Wang, P. and Li, X. (2020). Assessing the quality of information on wikipedia: A deep-learning approach.
- Warncke-Wang, M., Cosley, D., and Riedl, J. (2013). Tell me more: an actionable quality model for wikipedia. DOI: 10.1145/2491055.2491063.
- Wikipedia (2023a). Wikipédia:content assessment. Available at https://en.wikipedia.org/wiki/Wikipedia:Content_assessment. Accessed in 21st july, 2023.
- Wikipedia (2023b). Wikipedia:general disclaimer. Available at https://en.wikipedia.org/wiki/Wikipedia:General_disclaimer. Accessed in 21st july, 2023.
- Wikipedia (2023c). Wikipedia:size of wikipedia. Available at https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. Accessed in 21st july, 2023.