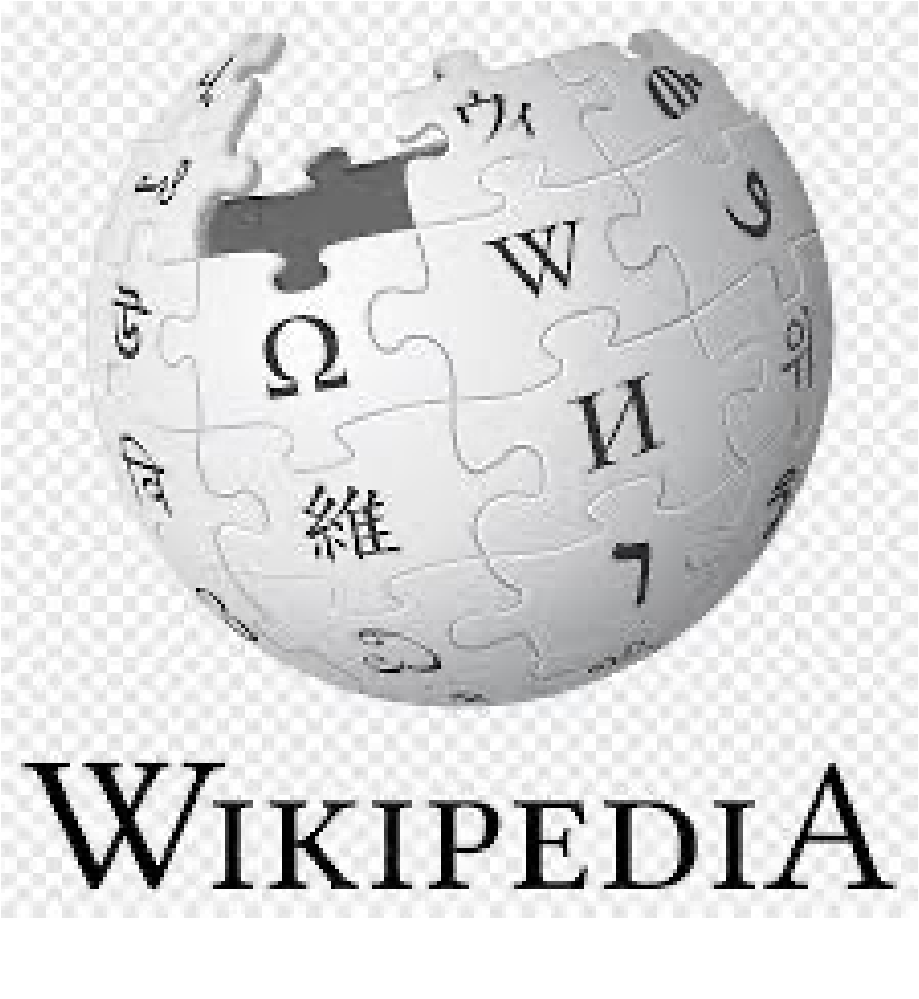
Wikipedia Article Quality Prediction

Final Report: DLSS

Nafiseh Tavakol, Kuon Ito, Lorenz, Rückert, Marius Helten

2025-08-21



```
Error in file(file, "rt"): cannot open the connection
```

```
Error in file(file, "rt"): cannot open the connection
```

Introduction

Wikipedia has become one of the most prominent sources of freely accessible knowledge worldwide, attracting millions of readers across domains such as history, politics, science, and popular culture (2020). Its openness to user contributions is both a strength—encouraging collaborative knowledge construction—and a weakness, since it may also lead to the creation of low-quality or biased articles (2019). While the platform employs a quality grading scheme ranging from Featured Articles (FA) to Stub-class entries to guide improvement efforts (2015),(2019), only a very small fraction of articles meet the highest standards, with just 0.09% classified as FA and 0.5% as Good Articles (2015).

To cope with this challenge, both manual and automated quality assessment approaches have been explored. Human volunteers and WikiProjects monitor articles, but the scale of Wikipedia and the subjectivity of evaluations make this process inconsistent and insufficient (2019). Early automated methods relied on handcrafted features of article text , while subsequent research applied machine learning techniques such as doc2vec (2014), BiLSTMs , and multimodal approaches incorporating images and metadata . The Neural Wikipedia Quality Monitor (NwQM) further demonstrated that integrating multiple modalities—including article text, metadata, and talk pages—substantially improves prediction performance (2021). Similarly, supervised classification frameworks have been developed using syntactic, stylistic, and editorial-history features, combined with labeled ground-truth datasets, to classify articles across the Wikipedia quality scale (2019).

Another crucial dimension of quality in Wikipedia is verifiability. The Verifiability policy requires that claims be supported by reliable sources, and unsourced material may be challenged with a {citation needed} tag (2019). Despite this, many articles contain unverified claims, and citation practices among editors are often ad hoc rather than systematic. As of 2019, over 350,000 articles contained at least one {citation needed} flag, but the actual number of unverified statements is likely far higher (2019). Understanding citation practices is therefore central to ensuring Wikipedia's reliability, since citations provide a strong signal of article quality and credibility.

Recent advances in graph representation learning have created new opportunities to model Wikipedia not only through its textual and editorial features, but also through its citation structures. Graph Neural Networks (GNNs) have proven highly effective at learning from graph-structured data, with applications in semi-supervised node classification and link prediction (2016). Traditionally, benchmarks for GNNs have relied on citation networks such as Cora, CiteSeer, and PubMed (2016). However, these datasets share similar structural properties and inconsistent training splits, limiting fair evaluation (2018). To address this, the Wiki-CS dataset was introduced as a Wikipedia-based benchmark with higher connectivity and structural diversity, offering a richer environment for testing GNN performance (2020).

Building on these insights, this project explores the potential of Wikipedia's citation networks as a foundation for predicting article quality ratings. By leveraging Graph Neural Networks, we aim to capture how signals of reliability and authority flow through citation structures. This network-oriented perspective complements prior text-based and multimodal approaches, addressing gaps in the literature and offering new insights into the relationship between verifiability, article interconnectedness, and collaborative knowledge quality.

Related work

Bassani and Viviani (2019) proposed a supervised classification approach for Wikipedia article quality, introducing 264 handcrafted features spanning text, review history, and network dimensions. They highlighted the importance of reliable ground truth construction, showing that inconsistencies between labeled and current article versions can reduce accuracy. Their experiments demonstrated that text features were the most effective, while network features contributed less, and that Gradient Boosting achieved the best performance, reaching 62% accuracy in multi-class classification.

Reddy et al. (2021) introduced NwQM, a neural framework for article quality assessment that integrates text, metadata, and images to build multimodal representations. Unlike earlier approaches relying primarily on structural or handcrafted features , their model demonstrated an 8% improvement over state-of-the-art methods, establishing the value of multimodal learning in quality prediction.

Redi et al. (2019) examined verifiability in Wikipedia by introducing a taxonomy of citation reasons and developing models to predict both citation need and citation purpose. Their findings emphasized that citations are critical for maintaining Wikipedia's reliability, with certain types of content—such as historical facts, statistics, or reported speech—being particularly likely to require verification.

Mernyei and Cangea (2020) proposed Wiki-CS, a benchmark dataset for evaluating Graph Neural Networks (GNNs). Unlike traditional citation network datasets such as Cora, CiteSeer, or PubMed, Wiki-CS offers higher connectivity and a richer structure, making it a more challenging and diverse testbed. It is primarily used for semi-supervised node classification and provides standardized splits, addressing reproducibility issues in GNN research.

Together, these studies highlight the progression from handcrafted features to multimodal learning and benchmark datasets, and they underscore the importance of verifiability and structured evaluation. Building on these insights, our project focuses on leveraging citation networks and GNNs to capture how quality signals propagate across interconnected articles, thereby offering a complementary network-based perspective on Wikipedia article quality assessment.

Data Analysis

Dataset Description

Features:

A dataset of 379,926 English Wikipedia articles was assembled to support the quality prediction task. The collection combines article-level features with editing history, allowing both structural and behavioral dimensions of quality to be examined. Articles span all major quality classes, from Stub to Featured Article (FA), which ensures that the dataset covers a broad range of writing styles, completeness levels, and editorial efforts. The design of this dataset is informed by previous research on Wikipedia quality. Earlier studies have focused on text features, structural elements, and editorial history (2019), multimodal features such as images and metadata (2021), and verifiability signals like citation practices (2019). In parallel, graph-based datasets such as Wiki-CS (2020) have highlighted the potential of network structures for modeling knowledge propagation. Drawing on these insights, the dataset integrates both content- and network-oriented variables to enable analysis from multiple perspectives. For each article, descriptive features such as page length, number of references, sections, templates, and infobox presence were recorded. Structural metadata such as categories, links, and depth in the category hierarchy were also included. In addition, editorial activity was carefully tracked: edit counts were separated into human and bot contributions, and further divided by registered, anonymous, and automated accounts. To capture recent editing patterns, all edit-related variables were restricted to the last two years. This ensures that the dataset reflects up-to-date collaborative dynamics rather than the full historical record.

Table 1: Variables grouped by category with their definitions

Category	Variable	Definition
Structure	num_categories	Number of categories assigned to the article.

Category	Variable	Definition
Structure	num_links	Total number of internal/external links.
Structure	page_length	Length of the article (characters).
Structure	num_references	Number of citations in the article.
Structure	num_sections	Number of sections.
Structure	num_templates	Number of templates used.
Structure	has_infobox_encoded	1 if an infobox exists, otherwise 0.
Structure	protection_status_encoded	Encoded protection level.
Style / Semantic	assessment_source_umap_1	UMAP dim 1 of assessment source.
Style / Semantic	assessment_source_umap_2	UMAP dim 2 of assessment source.
Style / Semantic	assessment_source_umap_3	UMAP dim 3 of assessment source.
Network	days_since_last_edit	Days since the last edit.
Network	edits_all_types	Total edits (last two years).
Network	edits_anonymous	Anonymous edits (last two years).
Network	edits_bot	Bot edits (last two years).
Network	edits_group_bot	Group-bot edits (last two years).
Network	edits_human	Human edits (last two years).
Network	edits_name_bot	Named-bot edits (last two years).
Network	edits_user	Registered-user edits (last two years).
Network	pageviews_Jul2023Jul2024	Pageviews from Jul 2023–Jul 2024.

Edges: ......

Target Variable:

Wikipedia articles are rated on an ordinal quality scale. In this project the following classes are used as the target: FA, FL, FM, A, GA, B, C, Start, Stub, List.

Table 2: Wikipedia quality assessment classes and their meaning

Class	Meaning
FA	Featured Article – highest quality, comprehensive and well-sourced
FL	Featured List – best-quality lists, complete and well-referenced
FM	Featured Media – high-quality non-textual media (images, videos, etc.)
A	Near-featured quality, but may need minor improvements
GA	Good Article – accurate, well-structured, but less comprehensive than FA

Class	Meaning
B	Mostly complete, but still lacking references or polish
C	Useful coverage, but incomplete or missing important details
Start	Basic coverage, underdeveloped but beyond stub level
Stub	Very short or incomplete article, minimal information
List	Articles in list format, assessed on completeness and structure

Data Collection:

For this project, four complementary data sources were used to capture different aspects of Wikipedia articles. The Wikipedia Dump provided raw text and structural information. The Pageviews API measured popularity and user attention. The Edit History API tracked editorial activity patterns, including user and bot edits. Finally, the Wikipedia API enabled article crawling and the construction of article networks. Together, these sources ensured a comprehensive dataset covering content, usage, editing behavior, and network structure.

Page IDs:

Article titles were retrieved from Wikipedia based on their page IDs. Since the MediaWiki API only allows up to 50 IDs per request, page IDs were split into chunks of 50. Each chunk was queried in parallel using ThreadPoolExecutor, and the resulting titles were mapped back to the original dataframe. The updated dataset, containing both pageid and title, was saved as a CSV file for further processing. Add more……

Text Features:

……

Edges:

…..

Article Titles:

A dedicated script retrieved article titles from Wikipedia by mapping page IDs to titles. Requests were executed in parallel to speed up the process, and results were stored for integration into downstream analyses.

Pageviews:

One year of pageview data (July 2023–July 2024) was collected for each article. Monthly user pageviews were retrieved from the Wikimedia REST API and aggregated into annual totals. To improve efficiency, requests were executed in parallel with ThreadPoolExecutor, and progress was tracked with tqdm. Results were incrementally written back to the dataset to avoid data loss in case of interruptions. The final dataset, articles_page_view.csv, contains each article with its total pageviews for the year.

Last Edit:

To obtain temporal metadata, last edit timestamps were collected using the REST API's /page/summary/{title} endpoint. Article titles from articles_titles.csv were queried with randomized delays (0.3–0.6s) and retry logic for HTTP 429 errors. Parallel requests (via four workers) accelerated collection. Errors and rate limits were logged, and results were stored in final_last_edit.csv, containing article titles and their last edit timestamps.

Edit Counts:

Wikipedia edit counts (July 2023–July 2025) were collected by editor type: registered users, anonymous users, group bots, and name bots. To comply with Wikimedia's API limits, randomized delays and proxy rotation were implemented when rate limits occurred. Edits were aggregated into totals and separated into human vs. bot contributions. Parallel processing with ThreadPoolExecutor improved efficiency, while periodic checkpoints ensured progress was saved. The output provides structured edit counts per article for subsequent analysis.

Article Target Feature Processing

This stage prepared the Wikipedia article dataset for downstream modeling by constructing target labels and engineering structured features.

Data Loading and Indexing:

The dataset was loaded using Polars, and an index was created containing article titles, page IDs, and internal numeric identifiers to support efficient lookup.

Target Construction

Each article was mapped to its Wikipedia quality class (FA, GA, B, etc.). From these labels, three target variables were derived:

- Target_QC_cat – a 10-level ordinal scale (List=0 … FA=9).
- Target_QC_aggcat – a 3-tier aggregate scale (Low, Mid, High).
- Target_QC_numlog – a log-transformed numeric variant to smooth scale differences.

Feature Encoding:

Several categorical and binary attributes were encoded:

- Protection status was label-encoded into integers (unprotected=0 … fully_protected=3).
- Infobox presence was represented as a binary indicator.
- Assessment source was one-hot encoded and reduced via UMAP into three dense components.

Feature Selection:

The final feature set integrated multiple dimensions of article characteristics:

- Content metrics – page length, number of sections, templates, references, categories, and links.
- Editorial activity – days since last edit and edit counts by different user types (e.g., human vs. bot).
- Popularity – annual pageviews (July 2023 – July 2024).

The processed dataset combines content structure, editing activity, and popularity indicators, forming a comprehensive feature representation suitable for graph construction and downstream machine learning models.

Graph Preparation for GNN Training

To prepare the dataset for Graph Neural Networks (GNNs), the graph-tool format (.gt) was converted into PyTorch Geometric Data objects. The conversion included extracting node features, edge features, and target labels. Node and edge properties of different types (numeric, boolean, categorical) were systematically encoded, with categorical variables handled through either label encoding or one-hot encoding.

To ensure stable model training, node features were standardized using multiple scaling techniques (StandardScaler, MinMaxScaler, RobustScaler, QuantileTransformer, PowerTransformer, and log-based robust scaling). Several variants of the dataset were generated, each reflecting a different scaling method.

Target variables were selected from node attributes with the prefix Target_, with Target_QC_aggcat used as the primary label for classification tasks. For each processed graph, two outputs were stored: a PyTorch tensor dataset (.pt file) for direct GNN training and a Parquet file containing scaled features and targets for inspection and debugging.

This preprocessing pipeline produced a clean, scalable dataset ready for downstream graph-based learning on Wikipedia article quality prediction.

Data Exploration

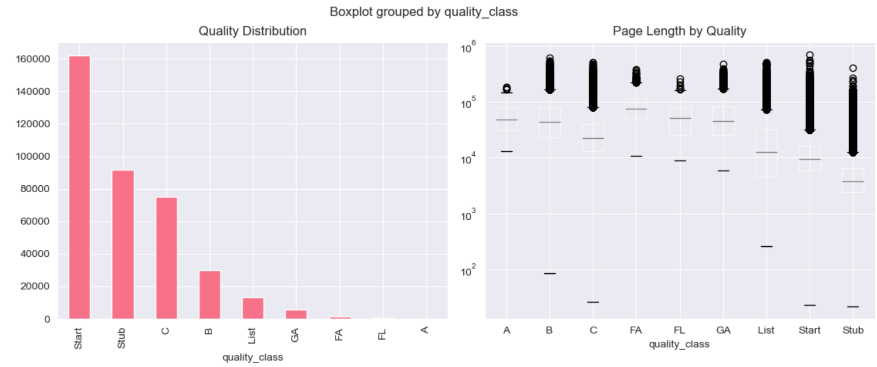
Network and Edit Structure

The dataset contains around 380,000 Wikipedia articles annotated with quality classes. The distribution is highly imbalanced, with the majority of articles falling into low-quality categories (Stub and Start), while only a small fraction belong to high-quality classes such as FA, GA, FL, and A. A clear progression is observed across the classes: higher-quality articles are substantially longer, as shown in the page length distributions, and they also tend to include more structural and citation features such as references, links, and sections. By contrast, Stub and Start articles remain short and sparsely referenced, reflecting limited editorial development. This confirms that structural richness and citation density strongly align with editorial quality.

Table 3: Quality classes and average structural metrics.

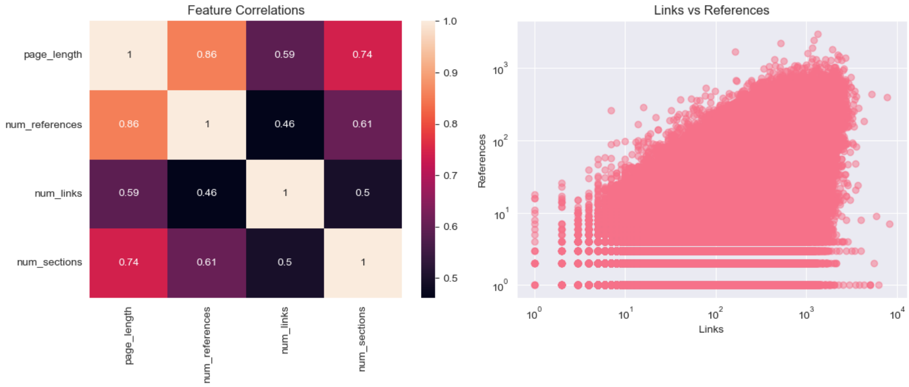
Quality Class	Count	Avg. Page Length	Avg. References	Avg. Links	Avg. Sections
A	113	60,096.51	97.15	386.11	17.62
B	29,768	61,135.93	96.63	422.76	21.02
C	74,983	33,138.79	47.86	285.05	14.69

Quality Class	Count	Avg. Page Length	Avg. References	Avg. Links	Avg. Sections
FA	1,582	89,048.49	142.50	514.55	21.72
FL	320	58,989.33	87.65	370.57	11.29
GA	5,934	64,000.73	115.96	395.99	17.88
List	13,161	29,490.59	29.69	368.69	13.75
Start	162,145	14,064.61	17.70	179.28	8.25
Stub	91,920	5,878.57	6.29	143.61	4.21



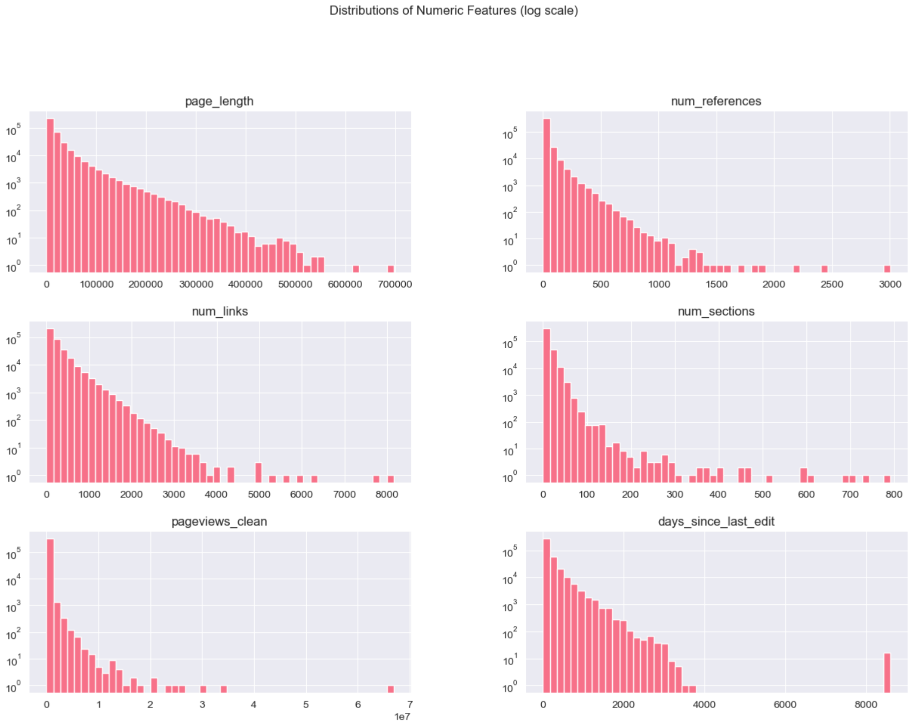
Feature Correlations and Relationships

The heatmap shows strong correlations between page length, references, and sections, with the highest between page length and references (0.86). This indicates that longer articles tend to be better structured and more referenced. Links are positively correlated but capture partly independent information. The scatter plot (log–log scale) of links vs. references confirms a positive trend: articles with more links usually also include more references, though variation exists. This shows that links and references represent complementary aspects of article richness.



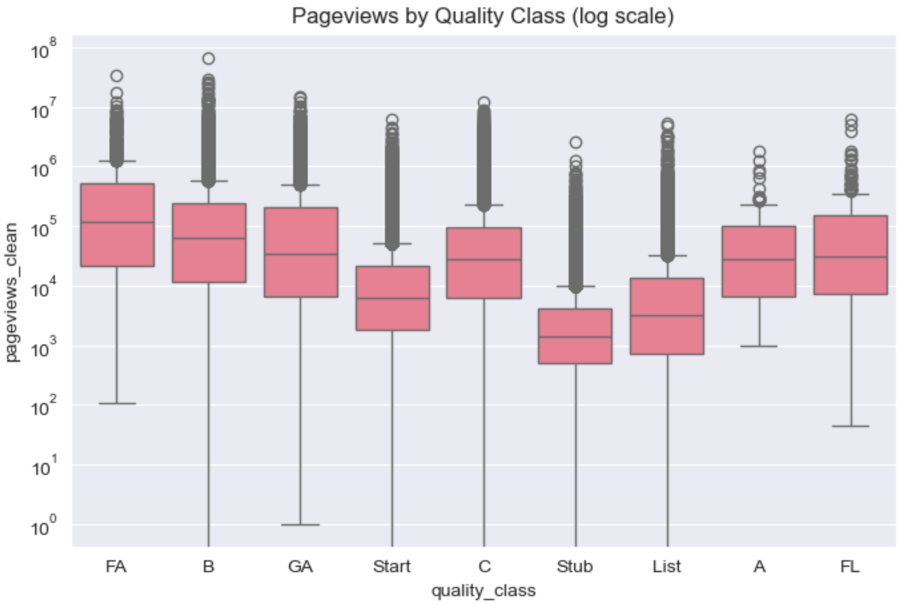
Feature Distributions

Key features such as page length, number of references, links, sections, pageviews, and recency of edits are highly skewed, with most articles clustered at the lower end and a few outliers extending to extreme values. This highlights the heterogeneity of Wikipedia, where a small subset of articles dominates in depth and attention.



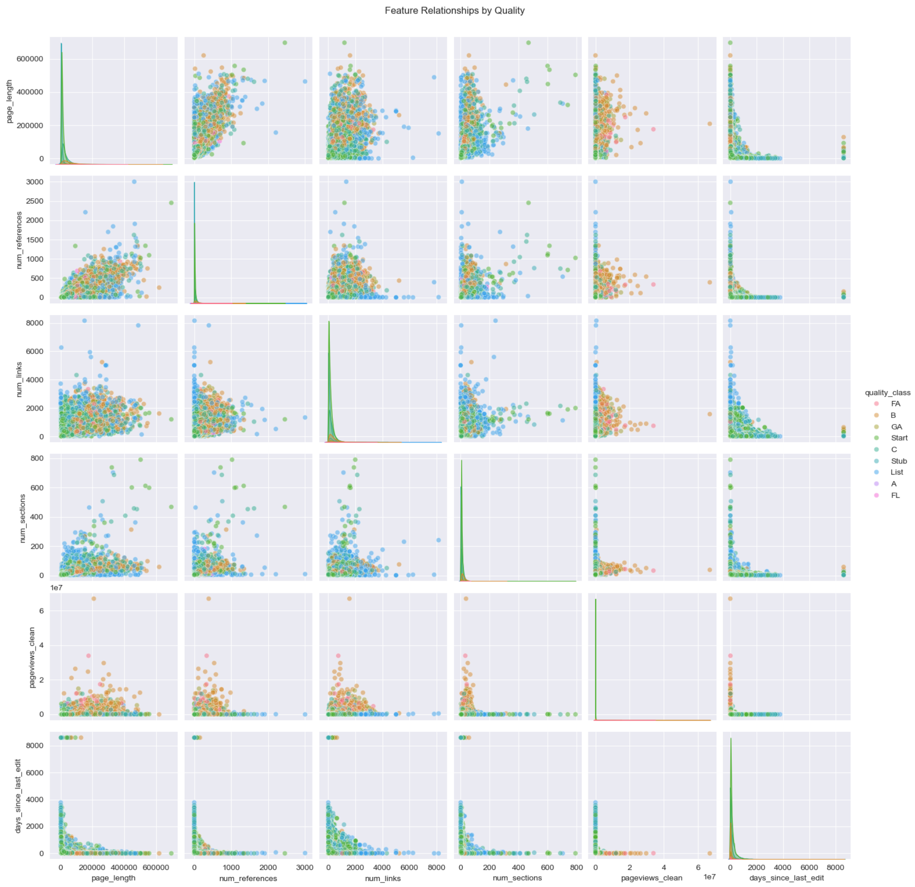
Pageviews by Quality

Pageviews vary widely across classes. While Featured Articles (FA) and Good Articles (GA) generally attract higher median views, many B-class and even lower-quality articles also reach high visibility. This suggests that popularity is not fully aligned with editorial quality, articles can be widely read even if their structural quality is limited.



Feature Relationships

Feature Relationships Pairwise feature comparisons show clear clustering by quality: high-quality articles combine length, references, links, and sections in consistent proportions, while low-quality articles remain compact across all dimensions. Pageviews and recency of edits add further variation but only partially align with quality, reinforcing that structural completeness and editorial effort are the strongest signals of quality.



Wikipedia Network

The network was obtained by a BFS-search, starting at a handful of seed articles.

Graph Description

- single CC
- directed network
- sparse network
- reasonably clustered
- pretty sizeable sample but by no means exhaustive for all of wikipedia

```
Error: object 'data_descriptive' not found
```

- Degree Distribution
- reasonably similar to powerlaw or lognormal - in any case heavy tailed (very few nodes with a lot of connections), pretty normal for many internet networks, rich-get richer effect

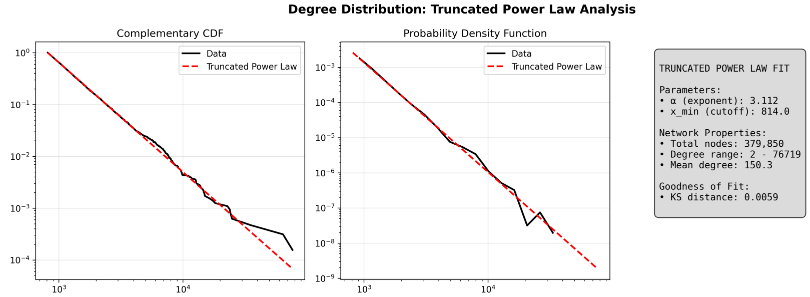


Figure 1: Wikipedia Graph - Degree Distribution and Power Law Analysis

- for none of the tested attributes assortativity could be measured, which is very surprising and also pretty disheartneing
- basically similar nodes are not neighbors which takes away one of the main theoretic assumptions

```
Error: object 'data_homophily' not found
```

- Transition Probabilties for all classes on graph

Graph based Features

- For one we used the conenctions in the network as part of the data that all the neural networks except for the MLP-Baseline were trained

- Centralities

  - pagerank
  - katz (paused beacause of redundancy)
  - betweenness
  - hub/authority
  - degree in
  - degree out
  - core number
  - clustering (local)

- CCDFs of network metrics across article categories

- Spectral Embedding
- Transition Matrix
- Modularity Matrix
- Share Reciprocity

Preprocessing

During the pre-processing we face two important challenges. The first was the highly imbalanced target variable with only a very small fraction of high-quality articles. The second challenge were the heavily skewed distributions, especially for the graph based features.

Regarding our target variable we tried switching from a classification to a regression problem. This allowed us to circumvent class counts by treating the ordinal categorical attributes as a numerical attribute. The heavily skewed distribution was log-transformed in order to obtain a less skewed distribution. After a brief evaluation, this approach proved to be flawed because the model just made average predictions. We concentrated on the classification approach. A first measure was to aggregated the categories into three ordinal classes from the initial count of nine. This improved class frequencies to a reasonable degree where training and predictions became possible.

The features had to be preprocessed since the numerical ranges were not uniform and fit for training. Here normalization could have solved the problem. However the distributions particularly for the network metrics are immensely positvely skewed. Different kinds such as standard, minmax, robust and robust-log scaling proved to be ineffective to generate reasonably spread distributions. The only approach that brought reasonable results was quantile scaling. Quantile scaling transforms data by mapping each value to its percentile rank, creating a uniform distribution where extreme outliers get compressed while preserving relative order.

For the network we remove nodes with total degree(k)$k \leq 1$, so the leaves of the network. # Methods

Graph-Neural-Network Models

Graph-Convolutional

Graph-Sage

Graph Attention

Training and Evaluation

Results

(Performance comparison tables, learning curves, confusion matrices)

Discussion and Conclusion

(Performance Comparison, Interpret the results in the context of social network theory, Key Findings and Implications)

References

Code and Data

- API´s
- Python Packages

Literature

- Citeable papers

Bassani, Elias, and Marco Viviani. 2019. "Quality of Wikipedia Articles: Analyzing Features and Building a Ground Truth for Supervised Classification." *Information Processing & Management* 56 (3): 975–88. https://doi.org/10.1016/j.ipm.2019.01.003.

Horta Ribeiro, Manoel, Kristina Gligoric, Maxime Peyrard, Florian Lemmerich, Markus Strohmaier, and Robert West. 2020. "Sudden Attention Shifts on Wikipedia Following COVID-19 Mobility Restrictions." *arXiv Preprint arXiv:2005.08505*. https://arxiv.org/abs/2005.08505.

Kipf, Thomas N., and Max Welling. 2016. "Semi-Supervised Classification with Graph Convolutional Networks." *arXiv Preprint arXiv:1609.02907*. https://arxiv.org/abs/1609.02907.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 1188–96.

Mernyei, Peter, and Cătălina Cangea. 2020. "Wiki-CS: A Wikipedia-Based Benchmark for Graph Neural Networks." *arXiv Preprint arXiv:2007.02901*. https://arxiv.org/abs/2007.02901.

Reddy, Bhanu Prakash, Sasi Bhushan, Soumya Sarkar, and Animesh Mukherjee. 2021. "NwQM: A Neural Quality Assessment Framework for Wikipedia." In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, 364–72. Association for Computing Machinery. https://doi.org/10.1145/3437963.3441754.

Redi, Miriam, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. "Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability." In *Proceedings of the World Wide Web Conference (WWW)*, 1567–78. ACM. https://doi.org/10.1145/3308558.3313618.

Shchur, Oleksandr, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. 2018. "Pitfalls of Graph Neural Network Evaluation." *arXiv Preprint arXiv:1811.05868*. https://arxiv.org/abs/1811.05868.

Warncke-Wang, Morten, Vladislav R. Ayukaev, Brent Hecht, and Loren G. Terveen. 2015. "The Success and Failure of Quality Improvement Projects in Peer Production Communities." In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, 743–56. ACM. https://doi.org/10.1145/2675133.2675241.

Yang, Zhilin, William W. Cohen, and Ruslan Salakhutdinov. 2016. "Revisiting Semi-Supervised Learning with Graph Embeddings." *arXiv Preprint arXiv:1603.08861*. https://arxiv.org/abs/1603.08861.