Wikipedia Article Quality Prediction

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

2025-08-26

Introduction

The Web enables anyone to read, publish, and share information at unprecedented speed and scale, greatly benefiting billions but also creating fertile ground for falsehoods (Kumar, West, and Leskovec 2016). Wikipedia, as one of the most widely used sources of free knowledge, faces credibility concerns due to hoaxes and the risk of low-quality or biased contributions (Horta Ribeiro et al. 2020; Kumar, West, and Leskovec 2016; Bassani and Viviani 2019). Although the platform employs a grading scheme from Featured Articles (FA) to Stubs, only a very small fraction of articles reach the highest quality levels, creating an imbalance that resembles anomaly detection, where rare but important cases must be identified (Warncke-Wang et al. 2015; Bassani and Viviani 2019). To address this, both manual and automated quality assessment methods have been explored. Human volunteers and WikiProjects monitor content, but scale and subjectivity limit their effectiveness. Automated approaches progressed from handcrafted textual features to machine learning models such as doc2vec (Le and Mikolov 2014), BiLSTMs, and multimodal systems that integrate images and metadata. Reddy et al. (Reddy et al. 2021) showed that multimodal learning substantially improves prediction, while Bassani and Viviani (Bassani and Viviani 2019) highlighted the challenges of reliable ground truth and found textual features more predictive than network ones. Verifiability is another key dimension: Redi et al. (Redi et al. 2019) introduced a taxonomy of citation reasons and showed that citation practices strongly signal credibility. Yet, as of 2019, more than 350,000 articles carried a tag, suggesting widespread unverified claims. Recent advances in Graph Neural Networks (GNNs) open new opportunities to model Wikipedia not only through text but also through its citation structures. Traditional citation benchmarks (Cora, CiteSeer, PubMed) suffer from limited diversity (Yang, Cohen, and Salakhutdinov 2016; Shchur et al. 2018), leading to the introduction of Wiki-CS, a richer Wikipedia-based dataset (Mernyei and Cangea 2020). Within this landscape, approaches can be divided into content-based, focusing on semantics and syntax, and context-based, emphasizing external signals such as social or citation networks (Monti et al. 2019).

Our project adopts a context-based perspective, leveraging both article relations (external references) and article structure (sections, citations, length). By applying GNNs, we aim to model how signals of reliability and authority propagate through these networks, while complementing them with additional structural features. This network-oriented approach avoids reliance on semantics or style, offering a generalizable and scalable framework for predicting Wikipedia article quality.

Data

Data Collection

This project combined four complementary data sources to capture different aspects of Wikipedia articles: the Wikipedia Dump (raw text and structure), the Pageviews API (popularity and user attention), the Edit History API (editorial activity patterns, including user and bot edits), and the Wikipedia API (article crawling and network construction). Article metadata was retrieved by mapping page IDs to titles via the MediaWiki API, which only supports up to 50 IDs per request; page IDs were split into batches of 50, queried in parallel with ThreadPoolExecutor, and merged back into the dataset before saving to CSV. One year of pageview data (July 2023–July 2024) was collected for each article from the Wikimedia REST API, aggregated into annual totals, and stored incrementally to prevent data loss; parallel requests and tqdm progress tracking ensured efficiency. Temporal metadata was added by retrieving last edit timestamps through the REST API's `/page/summary/{title}` endpoint, using randomized delays (0.3–0.6s), retry logic for HTTP 429 errors, and parallel workers to accelerate processing; results were saved to `final_last_edit.csv`. Editorial activity was captured for July 2023–July 2025, with edit counts broken down by registered users, anonymous users, group bots, and name bots, then aggregated into human vs. bot contributions. To respect API limits, randomized delays, proxy rotation, and periodic checkpoints were used, and data collection was parallelized for efficiency. Together, these steps produced a comprehensive dataset covering article text and structure, popularity, recency, editorial activity, and network relationships, providing a robust foundation for downstream analyses.

The network was obtained by a BFS-search starting at a handful of seed articles (**exact number at least**) and then expanding this seed through following internal links until a sufficient network size was reach. This gave a snowball sample of the whole of wikipedia. It has to be noted that this provides a biased sample of the network, that by no means is representative of all of wikipedia.

Dataset Analysis

Article Features

A dataset of 379,926 English Wikipedia articles was assembled to support the quality prediction task. The collection combines article-level features, structural metadata, and editing history, enabling both content-independent and behavioral dimensions of quality to be examined. Articles span the full range of Wikipedia's grading scheme, from Stub to Featured Article (FA), ensuring coverage of different writing styles, completeness levels, and editorial efforts. The design of this dataset is informed by prior research on text, structure, and verifiability (2019), (2021), (2019) as well as graph-based benchmarks such as Wiki-CS (2020). Drawing on these insights, the dataset integrates both article-level descriptors and network-oriented variables.

For each article, descriptive attributes include page length, number of references, number of sections, templates, infobox presence, and pageviews. Structural metadata records the number of categories, links, and depth in the category hierarchy. Editorial activity is tracked through detailed revision histories, separating human and bot edits and further distinguishing between registered, anonymous, and automated accounts. To reflect recent collaboration dynamics, edit-related variables were restricted to the past two years. Finally, additional context such as last edit timestamp, days since last edit, and protection status was included to capture recency and stability. This design results in a dataset that captures both structural and editorial signals, complementing traditional content-based features and enabling a multi-perspective analysis of Wikipedia article quality.

Table 1: Variables grouped by category with their definitions

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

Target Variable: Article Quality

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
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 Exploration

The dataset comprises ~380,000 Wikipedia articles labeled with quality classes, but the distribution is highly imbalanced: most articles fall into low-quality categories (Stub, Start), while only a small minority reach high-quality levels (FA, GA, FL, A). Quality progression is evident—higher-quality articles are much longer and include richer structural and citation features such as references, links, and sections. In contrast, Stub and Start articles remain short and sparsely referenced, reflecting limited editorial development. These patterns confirm that structural richness and citation density are closely associated 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
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

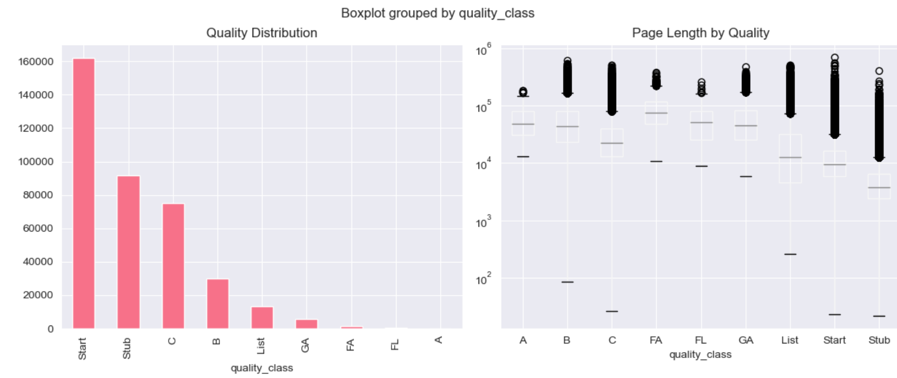


Figure 1: Wikipedia Graph - Degree Distribution and Power Law Analysis

The heatmap shows strong correlations among structural features, with the highest between page length and references (0.86), indicating that longer articles are usually better structured and more thoroughly referenced. Links are also positively correlated but provide partly independent information. A log–log scatter plot of links versus references confirms this trend: articles with more links often include more references, though variation remains, showing that links and references capture complementary aspects of article richness.

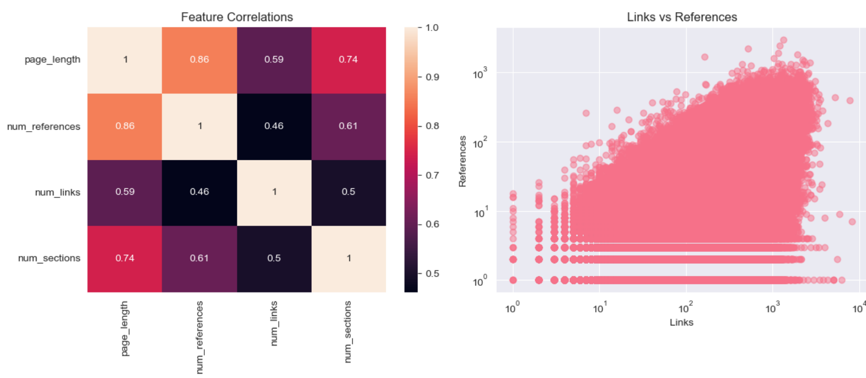


Figure 2: Wikipedia Graph - Degree Distribution and Power Law Analysis

In Feature Distributions plots, most articles cluster at the low end for page length, references, links, sections, pageviews, and recency of edits, with only a few outliers reaching extreme values—reflecting Wikipedia's heterogeneity, where a small subset dominates in depth and attention..

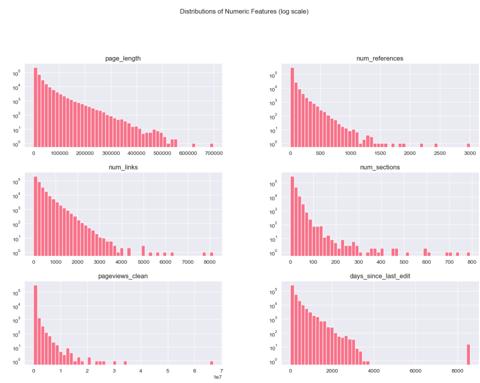


Figure 3: Wikipedia Graph - Degree Distribution and Power Law Analysis

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.

Figure 4: Wikipedia Graph - Degree Distribution and Power Law Analysis

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.

Network Description

As already mentioned we create the network from the sampled articles using the internal links of wikipedia articles. This way we obtained a directed network with a single weakly connected component. The network has a The network is very sparse (Density = 0.0002), which is no suprise for a network of this size. There is also some clustering in the network and the graph has a small diameter (approximated). This make sense since the articles were obtained by BFS and collection did not go further than 2-3 steps. Notable is the high share of reciprocal relations which shows that many articles link each other.

Table 4: Network Descriptive Metrics

Metric	Value
Nodes	379850.00000
Edges	28541137.00000
Density	0.00020
Reciprocity	0.46768
Global Clustering	0.11369
Pseudo-Diameter	7.00000

The degree distributions for the network are very long tailed, which is typical for many internet and citation networks. This can the consequence of the age of articles or some sort of preferential attachment or local redirection mechanism.
We see some difference between In- and Out-Degree. In-degrees have a much higher range, up to 75k, while Out-degrees are much smaller, up to 2,5k. This is because an article can be linked to many more articles than it can link itself. Interestingly the means are very similar. The power-law is a marginally better fit for the In-degrees (smaller KS-distance) and α_{In} is lower, but still above 2.

Figure 5: Degree Distribution with Power Law Fits

For none of the tested attributes strong assortativity was measured, with different implications. For the degree there is no homophily or heterophily between articles of similar or dissimilar, in- or out-degrees. Showing that the network is not forming structures along these properties.

Table 5: Network Assortativity

Assortativity	Value
In-Degree	0.0175
Out-Degree	0.0060
Categorical Quality	0.0124
Categorical Quality (Agg.)	0.0219
Numerical Quality	0.0292

For the Article-Quality the network is also not showing notable assortativity. There are only marginal increases by changing the encoding of the varibles, either by aggregating categorical varibles or switching to a numerical encoding. Higher values would have proved that articles form homogenous communities based on their quality. This already strongly discourages the hypothesis that Quality-signals propagate directly through channels in the network and only leaves the option of other node-properties being connected to certain qualities. It also already foreshadows why the direct network structure does not provide much help in classifying the article-nodes, since many GNNs rely on homophily by aggregating information from the neighborhood of nodes.

Network Features

To further enrich the dataset we use the wikipedia graph to create network based features for the article-nodes. The hope is that these will provide crucial additional information to help classifying them. For example it is conceiveable that certain article-qualities are associated with certain structural positions in the network.

Table 6: Network-based features extracted from Wikipedia graph.

Feature Category	Feature Name	Description
Degree Centrality	In-Degree	Number of incoming links to an article
Degree Centrality	Out-Degree	Number of outgoing links from an article
Local Structure	Clustering Coefficient	Measure of local clustering around a node
Path-based Centrality	Betweenness Centrality	Measure of how often a node lies on shortest paths
Core Structure	Coreness Centrality	Deepest k-core a node belongs to
Link Analysis	PageRank	Global importance based on link structure
Link Analysis	HITS Hub Score	Authority as a hub pointing to other authorities
Link Analysis	HITS Authority Score	Authority as a source pointed to by hubs
Reciprocity	Share of Reciprocal Relations	Proportion of bidirectional connections

Feature Category	Feature Name	Description
Spectral Features	Spectral Embedding (9D)	Low-dimensional representation using eigenvectors
Spectral Features	Spectral Modularity Row Sums	Row sum values from spectral modularity matrix
Probabilistic Features	Transition Probability Max	Maximum probability from transition matrix

In the following plots we have separated the articles by their quality. For consistency, the observed trend should preserve the quality-category order (HQ-MQ-LQ). The plotted values are pre-quantile-scaling as that would have greatly hampered their interpretability.

For In-Degree we see the very low share of very high in-degree articles, mainly in the MQ category. On the lower end of the in-degrees (< 5000) we see that HQ articles have a higher share of articles with increasing in-degree. For the Out-Degree this is much more visible, showing that higher quality articles tend to link to more articles.

The clustering shows that higher quality articles tend to be less clustered, so more functioning as hubs, inhabiting bridging positions, also when having more conenctions share of relations between neighbors rises much slower. Which is also supported by the relatively higher Betweenness centrality. Here the Coreness Centrality somehow speaks of a different picture, showing higher quality networks are more deeply embedded into the network.

For Page-Rank and both HITS centralities (Hubs and Authority) we see clearly, at least for the Hubs-Score, that higher Quality articles posses higher scores. Unfortunatly this becomes less clear for Authority and Pagerank scores, because of the width of the distributions, showing that a large majority of all articles posses very low scores, but the general trend (HQ>MQ>LQ) held. Here the sampling method might have produced these exceptionally high values, though this is also conformed by the other node-level measures.

A very interesting case is the share of reciprocal relations. Here we see for HQ and MQ higher shares of low reciprocity and lower shares of high reciprocity. Interestingly LQ articles show a very steady almost linear decline. Which also fits well with the In- and Out-degree results.

All these results point towards a qualitative difference in node-properties relating to article quality and thus should provide helpful information for classification.

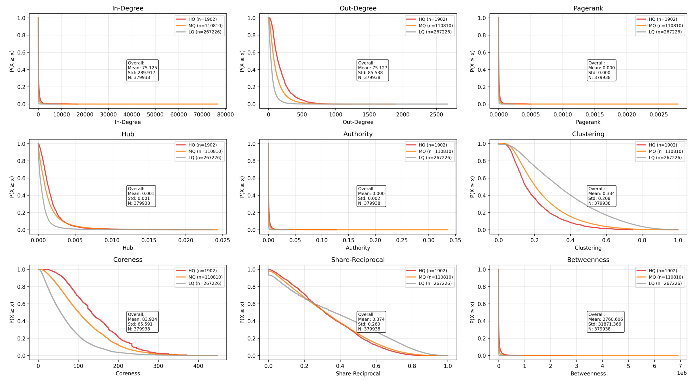


Figure 6: CCDF for Network Features of Wikipedia Graph

Furthermore we created a 9-Dimensional Spectral Embedding for the graph and used it as features for the dataset. Spectral embedding maps nodes to a low-dimensional space using eigenvectors of graph matrices (like the Laplacian), where the geometric distances preserve the graph's structural relationships. The embedding for the 2-D plot was chosen by the largest eigenvalue gap. There is some notable separation over the space, notably along 3-lines. Here probably also the sample structure played a major role in shaping this feature. In the plot we can also see that certain areas are more densely populated by either green or blue, which could be taken as a sign that certain article qualities inhabit a different structural position in the network. This would also support our earlier findings for the other network-features.

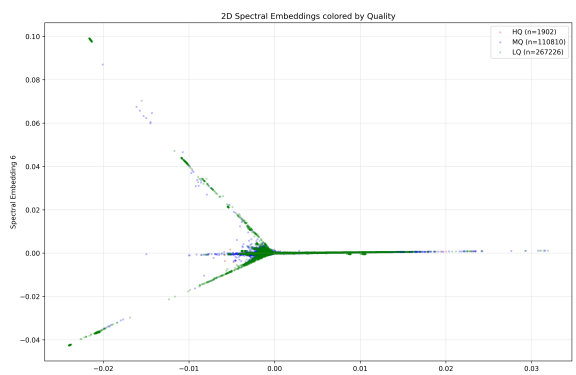


Figure 7: 2D Spectral Node Embedding produce by Wikipedia Graph Sample

We also tried spetral modularity based features and kept the row sum values and for transition probability based features we kept the maximum probability.

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. These have to be adressed to get

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 then concentrated on the classification approach. A first measure was to aggregated the categories into three ordinal classes (LQ, MQ and HQ) from the initial count of nine. This improved class frequencies to a reasonable degree where training and classifications became possible.

The features had to be preprocessed since the numerical ranges for many features, especially the network features, were very concentrated to a small range and thus not fit for training. Here normalization proved to be insufficient. The distributions particularly for the network metrics are immensely positvely skewed. Different scaling methods 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.

From the wikipedia graphs we removed nodes with total $degree(k) \leq 1$.

Article Target Feature Processing

The dataset was prepared for modeling by constructing target labels and encoding structured features. Using Polars, articles were indexed by title, page ID, and numeric identifiers for efficient lookup. Each article was mapped to its Wikipedia quality class (FA, GA, B, etc.), from which three target variables were derived: a 10-level ordinal scale (`Target_QC_cat`), a 3-tier aggregate scale (`Target_QC_aggcat`), and a log-transformed numeric variant (`Target_QC_numlog`). Categorical and binary attributes were encoded, including protection status (integer labels), infobox presence (binary), and assessment source (one-hot, then reduced with UMAP). The final feature set integrated content metrics (page length, sections, templates, references, categories, links), editorial activity (days since last edit, human vs. bot edits), and popularity (annual pageviews, July 2023–July 2024). Together, these features capture structural, editorial, and popularity dimensions of Wikipedia articles, providing a comprehensive representation for graph construction and machine learning models.

Graph Preparation for GNN Training

The Wikipedia dataset was converted from graph-tool format into PyTorch Geometric Data objects, with node features, edge features, and target labels systematically encoded (numeric, boolean, and categorical via label or one-hot encoding). To stabilize model training, node features were standardized using multiple scaling techniques (StandardScaler, MinMaxScaler, RobustScaler, QuantileTransformer, PowerTransformer, and log-based scaling), producing several dataset variants. Target variables were derived from `Target_` attributes, with `Target_QC_aggcat` used as the primary classification label. Each processed graph was stored in two forms: a PyTorch tensor dataset (`.pt`) for GNN training and a Parquet file for feature inspection and debugging. This pipeline yielded clean, scalable graph data suitable for downstream learning on Wikipedia article quality prediction.

Methods

Benchmarks

Machine-Learning

Multi-Layer-Perceptron

Graph-Neural-Network Models

- Hyperparameter Search

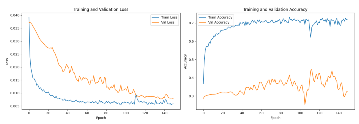
Graph-Convolutional

- Improved GNN
- Residual GCN

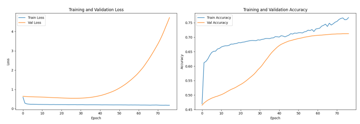
Graph-Sage

Graph Attention

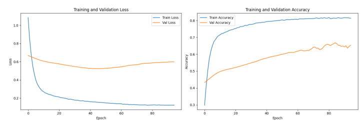
Training and Evaluation



(a) GCN



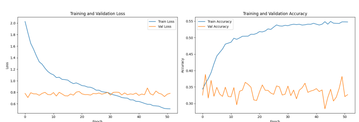
(b) Residual GCN



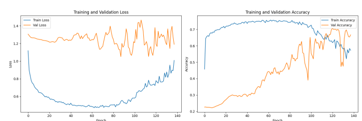
(c) Residual Sage

Figure 8: Test and Validation Loss and Accuracy during Training I

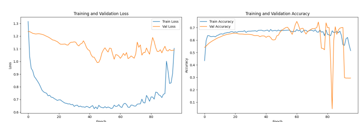
<-



(a) GAT



(b) MLP Network- and Article-Features



(c) MLP Article-Features

Figure 9: Test and Validation Loss and Accuracy during Training II

Results

(Performance comparison tables, learning curves, confusion matrices)

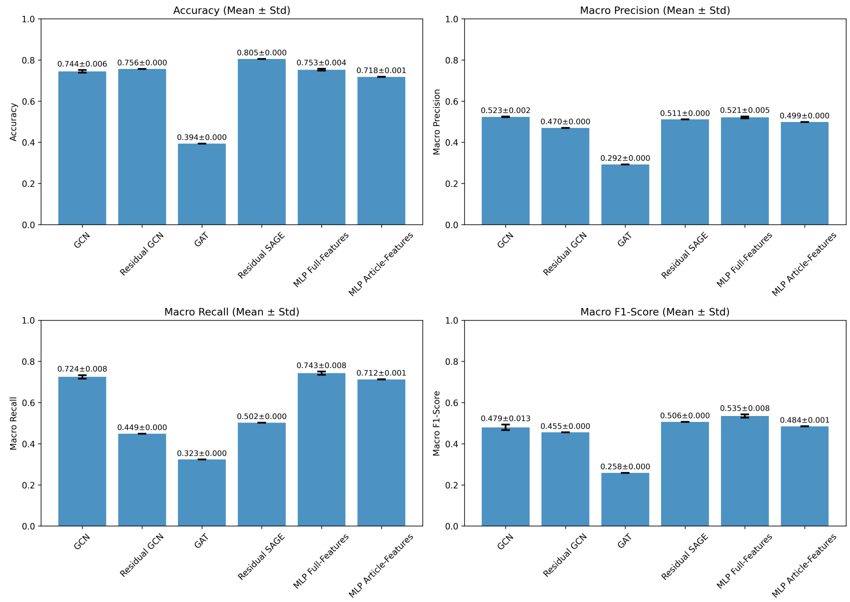


Figure 10: Evaluation Metrics for NN-Models

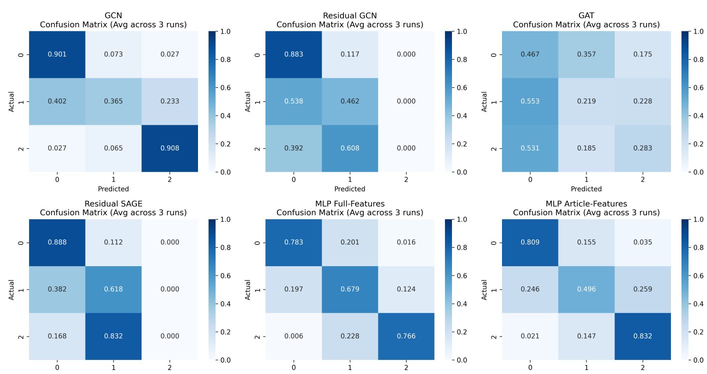


Figure 11: Confusion Matrices for NN-Models

Discussion and Conclusion

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

- Sampling method biased
- more extensive testing of architectures and parameters

- missing assortativity/homophily -> no direct quality propagation (providing network structure did not meaningfully improve GNNs over MLP)
- some hints for network structure and position being related article quality (MLP performed better), which by a few percentage points is already remarkable, as these have to be stacked in order to provide meaningful improvements

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.

Kumar, Srijan, Robert West, and Jure Leskovec. 2016. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes." In *Proceedings of the 25th International Conference on World Wide Web*, 591–602. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883085.

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.

Monti, Federico, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. "Fake News Detection on Social Media Using Geometric Deep Learning." arXiv. https://doi.org/10.48550/arXiv.1902.06673.

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.