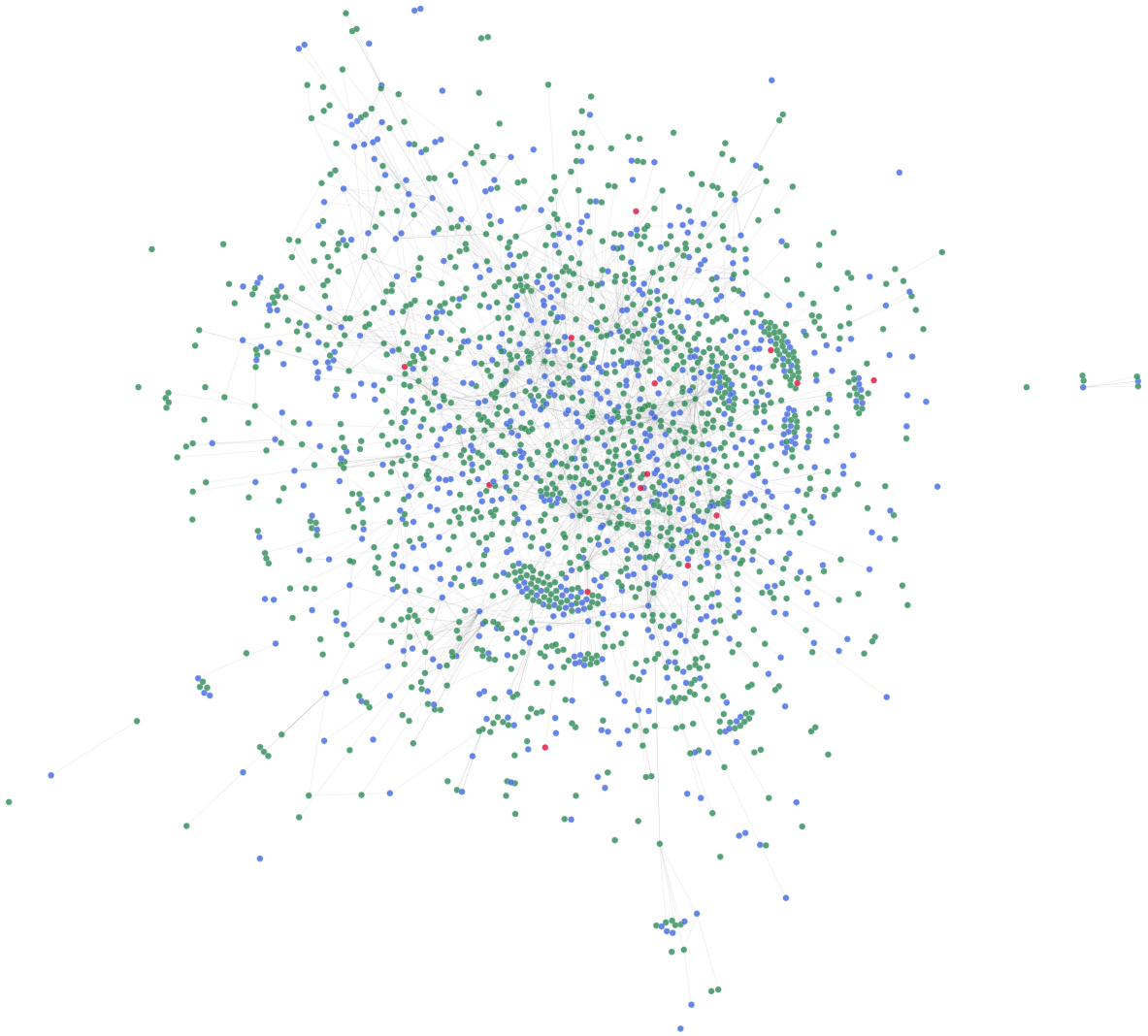
# Wikipedia Article Quality Prediction

Nafiseh Tavakol, Kuon Ito, Lorenz, Rückert, Marius Helten
https://github.com/Ari-manius/DLSS-WAQP

2025-08-30



## 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 (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 (2020; 2016; 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 (2015; 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 (2014), BiLSTMs, and multimodal systems that integrate images and metadata. Reddy et al. (2021) showed that multimodal learning substantially improves prediction, while 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. (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 (2016; 2018), leading to the introduction of Wiki-CS, a richer Wikipedia-based dataset (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 (2019).

Our project adopts a context-based approach, an alternative to text-driven methods, by leveraging article relations (internal links) and article structure (sections, citations, length) (2020). A central question here is whether Wikipedia's internal links function similar to academic citations, where high-quality papers often cite other influential works (2020). Unlike scholarly references, which carry signals of reliability and authority, Wikipedia's internal links are often added freely whenever a related article exists. Many wikilinks are automatically generated or inserted via templates (infoboxes, navigation boxes, etc.) (2020). By applying GNNs, we aim to explore to what extent signals of reliability and authority can propagate through these networks and how such context-based signals may complement structural features in predicting article quality.

## Data

### Data Collection

This project combined three complementary data sources to capture different aspects of Wikipedia articles: 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. 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 usage and periodic checkpoints were used; 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 with 4 seed articles (Influenza, Serena Williams, French Revolution, Quantum mechanics) and then expanding this seed through following internal links until a sufficient network size was reached. Each seed-article belongs to one of the four largest Wiki-Projects (Serena Williams → WikiProject Biography (largest), Influenza → WikiProject Medicine, French Revolution → WikiProject History,

Quantum mechanics → WikiProject Physics). This was done to get a solid variety of article data. Only the internal links pointing to articles in the sample were kept. 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. A better approach would have been to follow the one outlined by (2019).

**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 1: 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 |

### *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 2: 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. |

| Category | Variable | Definition |
|---|---|---|
| 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. |

**Article-Feature Analysis** 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.

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.

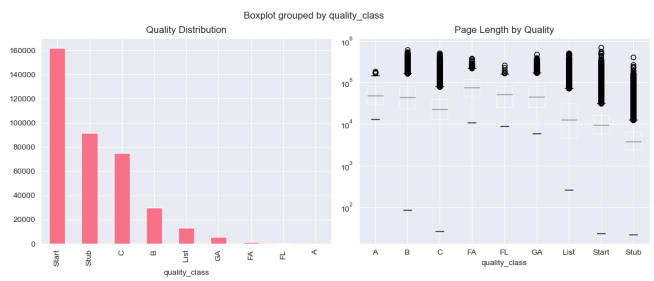



Figure 1: Distribution of Wikipedia article quality classes

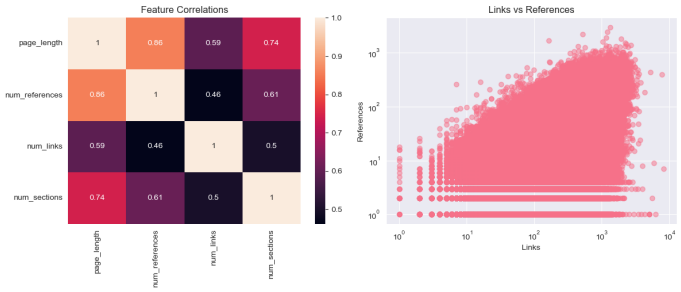



Figure 2: Correlation heatmap and link–reference scatter

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.

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 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.

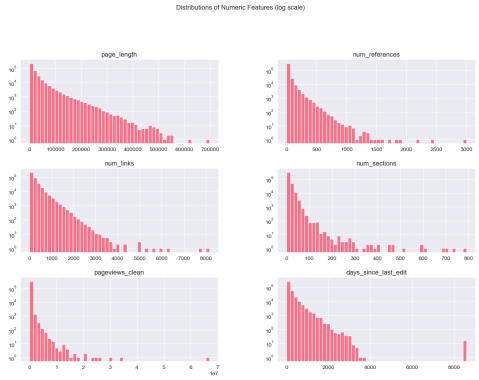



Figure 3: Distribution of numerical features

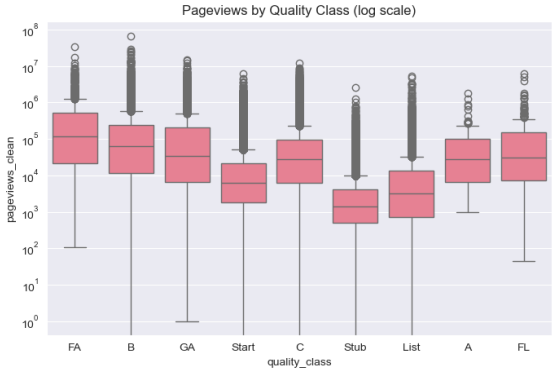



Figure 4: Pageviews distribution by article quality class

**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 is very sparse (Density = 0.0002), which is no surprise 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 4 steps. Notable is the high share of reciprocal relations which shows that many articles link each other.

Table 3: Network structure extracted from graph analysis

| Feature Category | Feature Name | Value |
|---|---|---|
| Basic Structure | Nodes | 3.798500e+05 |
| Basic Structure | Edges | 2.854114e+07 |
| Basic Structure | Density | 1.978000e-04 |
| Connectivity | Reciprocity | 4.676791e-01 |
| Local Structure | Global Clustering | 1.136937e-01 |
| Distance | Pseudo-Diameter | 7.000000e+00 |

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 $\alpha_{In}$ is lower, but still above 2.

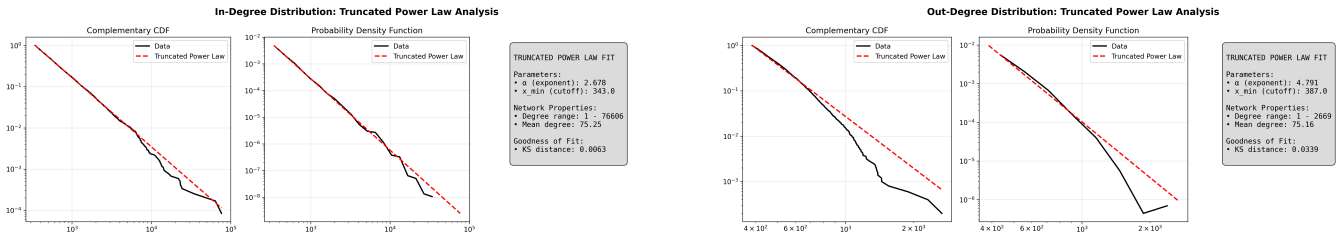



Figure 5: Degree Distributions with Power Law Fits

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 4: Assortativity features from network analysis

| Feature Name | Value |
|---|---|
| In-Degree | 0.0175411 |
| Out-Degree | 0.0059696 |
| Categorical Quality | 0.0124389 |
| Categorical Quality (Agg.) | 0.0219047 |
| Numerical Quality | 0.0291918 |

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 5: Network-based features extracted from Wikipedia Graph

| Feature Category | Feature Name |
|---|---|
| Degree Centrality | In-Degree |
| Degree Centrality | Out-Degree |
| Local Structure | Clustering Coefficient |
| Path-based | Betweenness Centrality |
| Core Structure | Coreness Centrality |
| Link Analysis | PageRank |
| Link Analysis | HITS Hub Score |
| Link Analysis | HITS Authority Score |
| Reciprocity | Share of Reciprocal Relations |
| Spectral Features | Spectral Embedding (9D) |
| Spectral Features | Spectral Modularity Row Sums |
| Probabilistic Features | Transition Probability Max |

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 conections 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.

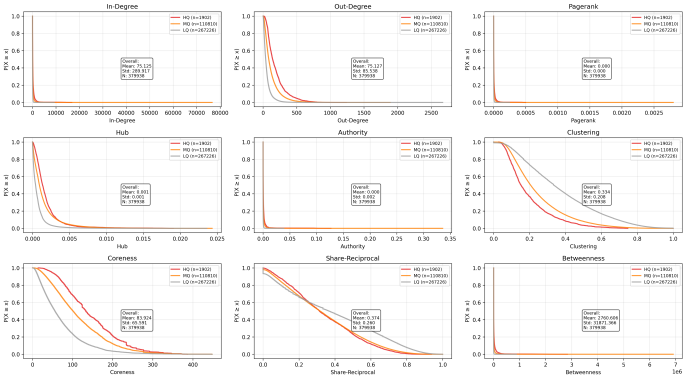



Figure 6: Complementary cumulative distribution functions for network features by quality category

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.

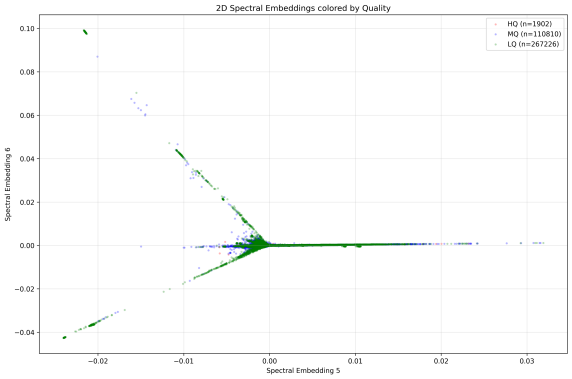



Figure 7: 2D spectral embedding of Wikipedia articles colored by quality

### *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

The dataset was prepared for modeling by constructing target labels and encoding structured features. 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`) (High Quality (HQ) > Medium Quality (MQ) > Low Quality (LQ) ), and a log-transformed numeric variant (`Target_QC_numlog`).

Categorical and binary article-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 information for the models.

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 (n=500) 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 $degree(k) \leq 1$.

An extra dataset with only article-features was created to compare if using the network features provided helpful information.

## *Methods*

### *Benchmarks*

To assess the added value of our graph-based models, we implemented several benchmarks: a Random Forest classifier, the ORES quality prediction service, and two Multilayer Perceptrons (MLPs).

**Random Forest** We train on the preprocessed tabular dataset (scaled_data_quantile_Target_QC_aggcat.parquet) with the aggregated three-class quality target. The feature table combines article-level attributes (e.g. length/structure/citations) and network-derived measures (e.g., centrality/degree-style features). For the Random Forest we use a fixed configuration (e.g. n_estimators=200, constrained depth, max_features='sqrt', bootstrap), fit on the training data (with validation used during development), and report accuracy, precision, recall, F1, and confusion matrices on the test set. We also assess permutation importances to gauge feature contributions.

**ORES (Objective Revision Evaluation Service)** We query the Wikimedia ORES article-quality model for the same test articles, using their revision IDs. ORES six-way predictions are mapped to the project's three aggregated classes for a like-for-like comparison. We then compute the same metrics as above and store detailed predictions for inspection. This positions our models against the operational standard in the Wikipedia ecosystem.

**Multilayer Perceptrons (MLPs)**

- Full Features
  The MLP is trained on the full set of available features, combining intrinsic article-based attributes with features derived from the network structure. The optimized model uses a hidden dimension of 256, two layers, a dropout rate of approximately 0.18, a learning rate of 0.00775, and a weight decay of 1.16×10^-4. Training employed a weighted cross-entropy loss along with boosted oversampling at a factor of two. Under these conditions, the validation score reached 0.810.

- Non-network Features
  The MLP is trained exclusively on article-level attributes, excluding all graph-derived information. The optimized hyperparameters specify a hidden dimension of 128, three layers, a very small dropout rate of about 0.014, a learning rate of 0.00144, and a weight decay of 2.46×10^-4. To address class imbalance, class-balanced focal loss was applied with  = 0.9999 and  = 2.55. In addition, boosted oversampling was used, increasing minority representation by a factor of four with a minimum boost of approximately five. This setup achieved a validation score of 0.805.

### *Graph-Neural-Network Models*

#### Graph-Convolutional

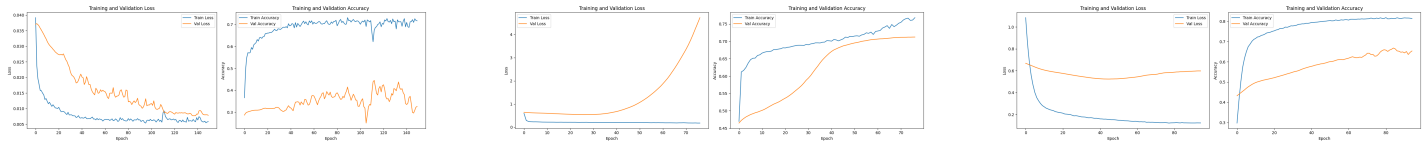
- Improved GNN
  The Improved GNN integrates multiple architectural enhancements to strengthen learning on imbalanced graph data. The model begins with input preprocessing using LayerNorm and a linear projection into the hidden dimension. It applies three layers of graph convolutions, configurable as either GraphSAGE or GCN, each followed by LayerNorm, GELU activation, and adaptive dropout that decreases in later layers for stability. Weighted residual connections are employed, with residual strength gradually increasing in deeper layers to maintain gradient flow and feature reuse. The final stage includes an output normalization and a projection layer for prediction. Optimization is performed with a learning rate of 0.0173 and weight decay of 4.2×10^-6. The training objective leverages Class-Balanced Focal Loss with parameters  = 0.9993 and  = 3.13, effectively addressing class imbalance. This is further reinforced by boosted oversampling with a minimum class boost of approximately 4.84 and a min_samples_factor of 4.
- Residual GCN
  This model employs a residual Graph Convolutional Network (GCN) architecture enhanced with modern design choices for stability and improved performance on imbalanced graph data. The input features are projected into a hidden dimension using a linear transformation followed by GELU activation. Two GCNConv layers are applied sequentially, each followed by LayerNorm, GELU activation, and dropout. Residual connections are incorporated with a weighting factor of  = 0.9, which prioritizes newly transformed representations while still retaining prior information to ensure gradient stability. A final linear projection produces the classification outputs. Optimization is performed with AdamW using a learning rate of 0.00173 and a weight decay of 2.95×10^-5. To handle class imbalance, Focal Loss is applied with  = 1.11 and  = 3.95, together with a boosted oversampling strategy that increases minority class representation by a factor of four.

**Graph-Sage** The model is based on a residual version of GraphSAGE, designed to enhance stability and performance on imbalanced graph data. The architecture begins with an input projection that embeds the raw node features into a hidden dimension, followed by two SAGEConv layers. Each layer applies LayerNorm, GELU activation, and dropout to improve expressiveness and regularization. Residual connections are incorporated with a weighted coefficient  = 0.85, allowing a balance between new transformed features and the identity mapping from the previous layer. A final linear layer produces the output predictions. Optimization uses AdamW with a learning rate of 0.00128 and weight decay of 0.00729. To address class imbalance, Focal Loss is applied with  = 1.00 and  = 3.98, combined with a balanced oversampling strategy that increases minority class representation by a factor of four.

**Graph Attention** It begins with the input features and passes them through three stacked GATConv layers, each followed by LayerNorm to stabilize training. The attention mechanism employs two heads, and the final layer averages their outputs to produce predictions. ELU activations and a dropout rate of approximately 0.64 are applied to enhance generalization and reduce overfitting. Training is carried out with AdamW optimization, using a learning rate of 0.00043 and a weight decay of 7.5×10^-5. To address class imbalance, Focal Loss is employed with parameters  = 2.83 and  = 3.78, while boosted oversampling is applied to increase minority class representation with a factor of two.

### *Training and Evaluation*

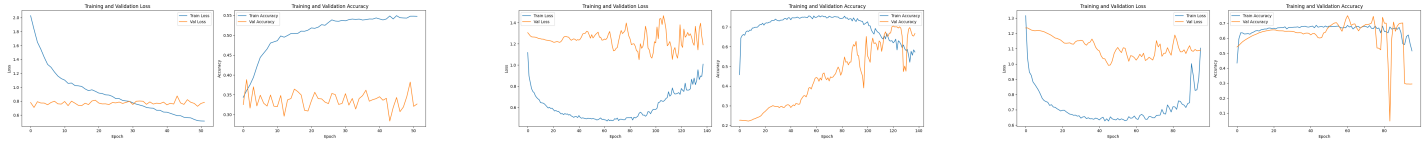
A range of different models was chosen for evaluation. With these candidates we performed a hyperparameter optimization over a limited search space. For each model 16 variants were created. Then started the finals runs where we used 3-fold cross-validation on different splits for each model.



(a) GCN (b) Residual GCN (c) Residual Sage

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

The dataset was partitioned using two different strategies depending on the experimental phase. For hyperparameter optimization, a 70/15/15 (train/validation/test) split was employed on a single random partition. For final model evaluation, we used an 80/10/10 split across three different random seeds (42, 142, 242) to assess model stability. This approach follows graph-aware splitting principles, preserving the connectivity structure essential for GNN performance.



(a) GAT (b) MLP Network- and Article-Features (c) MLP Article-Features

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

Graph neural networks (GCN, ResidualGCN, ResidualSAGE, GAT) show varying convergence patterns, with some models achieving faster validation accuracy improvements than others. There seems to be a lot of overfitting to training data, especially for the GAT where we dont see much change in validation loss and accuracy, while the training scores are increasing.

**Hyperparameter Optimization** Systematic hyperparameter optimization was conducted using Optuna with Tree-structured Parzen Estimator (TPE) sampling. The search space encompassed universal parameters including dropout rates (0.01-0.7), weight decay (1e-6 to 1e-2 on log scale), and learning rates (1e-4 to 1e-1 for most models, 1e-4 to 1e-2 for GAT). Model architecture parameters included hidden dimensions selected from [32, 64, 128, 256] and layer counts ranging from 2-5 (2-3 for GAT due to stability concerns). Model-specific parameters such as GAT attention heads [2,4,8,16] and loss function selection were also optimized.

The optimization objective combined 60% minority class F1-score with 40% overall accuracy, emphasizing performance on underrepresented high-quality articles while maintaining overall predictive capability. Each model underwent 16 trial, with 40 epochs per trial during optimization and 80-200 epochs for final evaluations.

**Model Training and Validation** Given the severe class imbalance in Wikipedia article quality ratings, multiple strategies were implemented to address this challenge. We experimented with specialized loss functions including Focal Loss ( =1.0-3.0,  =2.0-4.0), Class-Balanced Focal Loss ( =0.999-0.9999), and Weighted Cross-Entropy.

Additionally oversampling strategies were employed: “balanced” sampling ensuring equal representation per class, and “boosted” sampling with aggressive minority class enhancement (2.0-5.0x factors). To maintain computational feasibility, memory-safe sampling was limited to 20,000 samples maximum.

All models employed early stopping with patience values between 25-50 epochs and minimum delta thresholds of 1e-4 to prevent overfitting. Optimization utilized the Adam optimizer with OneCycleLR scheduling, featuring co-

sine annealing and 10% warmup periods. Regularization techniques included gradient clipping (max_norm=1.0) and model-specific dropout rates. Because of memory limitation, GraphSAINT subgraph sampling was implemented with batch sizes of 2048-8192, walk lengths of 2, and 5-8 sampling steps per epoch.

Model performance was assessed using standard classification metrics including accuracy, precision, recall, and F1-scores (macro because of class imbalance). Special attention was paid to minority class performance, given the practical importance of correctly identifying high-quality articles. Cross-validation results report mean performance with standard deviations across runs, providing confidence intervals for model comparisons. The MLP baselines, both with and without network features, provide important performance benchmarks for assessing the value of graph structure.

## Results

Our evaluation covered multiple GNN architectures, traditional ML baselines, and the ORES system. The cross-validation results show clear performance differences across approaches.

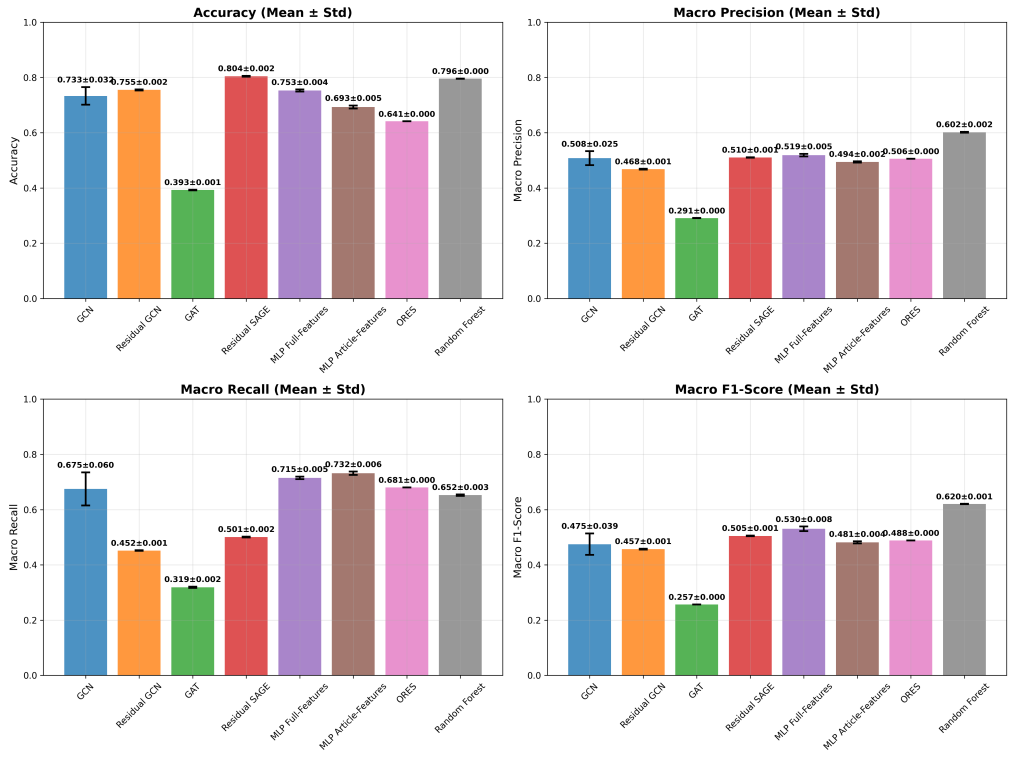



Figure 10: Cross-validation results for all models with error bars showing standard deviation across runs

Some models showed zero variance across runs (Residual GCN, GAT, Residual SAGE), while others had healthy variance around 0.4-0.6%. The zero-variance cases might indicate very stable convergence or potential issues with local optima.

Class imbalance handling varied significantly. Random Forest achieved the best macro F1-score (62.04%), while neural approaches clustered around 48-53%. This suggests the tree-based approach handles minority classes better than the neural networks, despite the various loss functions and sampling strategies we tried. Residual SAGE's success over other GNNs makes sense given our network analysis. Its sampling approach works better in sparse, heterophilic networks where traditional message-passing assumptions break down. Rather than assuming similar nodes cluster together, SAGE can aggregate from diverse neighborhoods, which fits Wikipedia's link structure better.

### Model Performance

Residual GraphSAGE achieved the highest accuracy at 80.50%, followed closely by Random Forest at 79.57%. The GNN approaches showed mixed results - while Residual GCN reached 75.62% and standard GCN got 74.45%,

GAT performed poorly at only 39.39%. This GAT failure likely stems from the networks lack of homophily, which we observed in our assortativity analysis. GAT's attention mechanism needs meaningful neighbor relationships, but our network structure doesn't provide this.

Comparing the MLP variants and investigating the importance of features used by the Random Forest reveals that network features matter: the full-feature model (75.26%) outperformed the article-only version (71.80%) by 3.46 percentage points. All our models beat the ORES baseline (64.12%) by substantial margins.

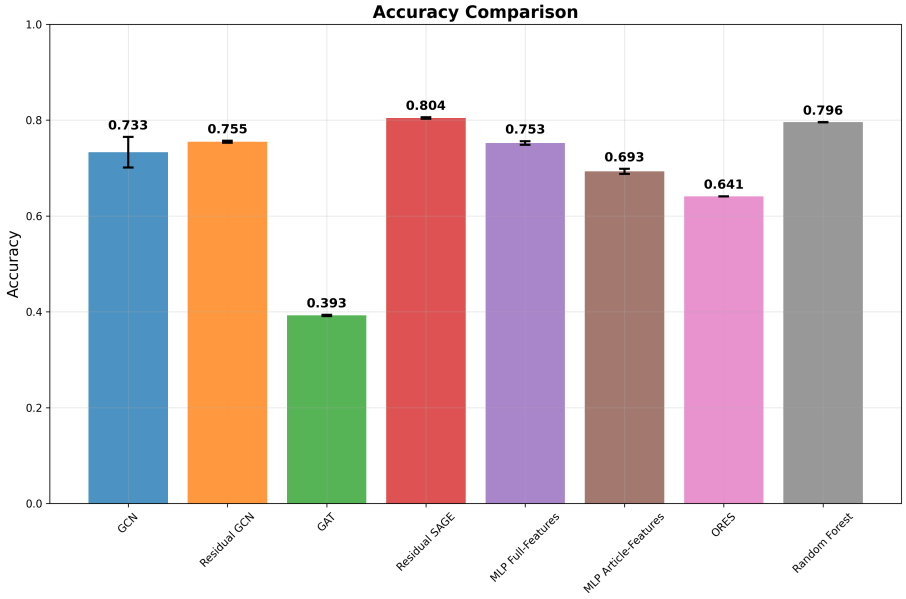



Figure 11: Final accuracy comparison across all evaluated models

### *Classification Patterns Across Model Types*

The confusion matrices show clear differences in how the models handle the three quality classes. The Random Forest stands out with the most balanced results across Low, Medium, and High quality articles, which is reflected in its stronger macro F1-score.

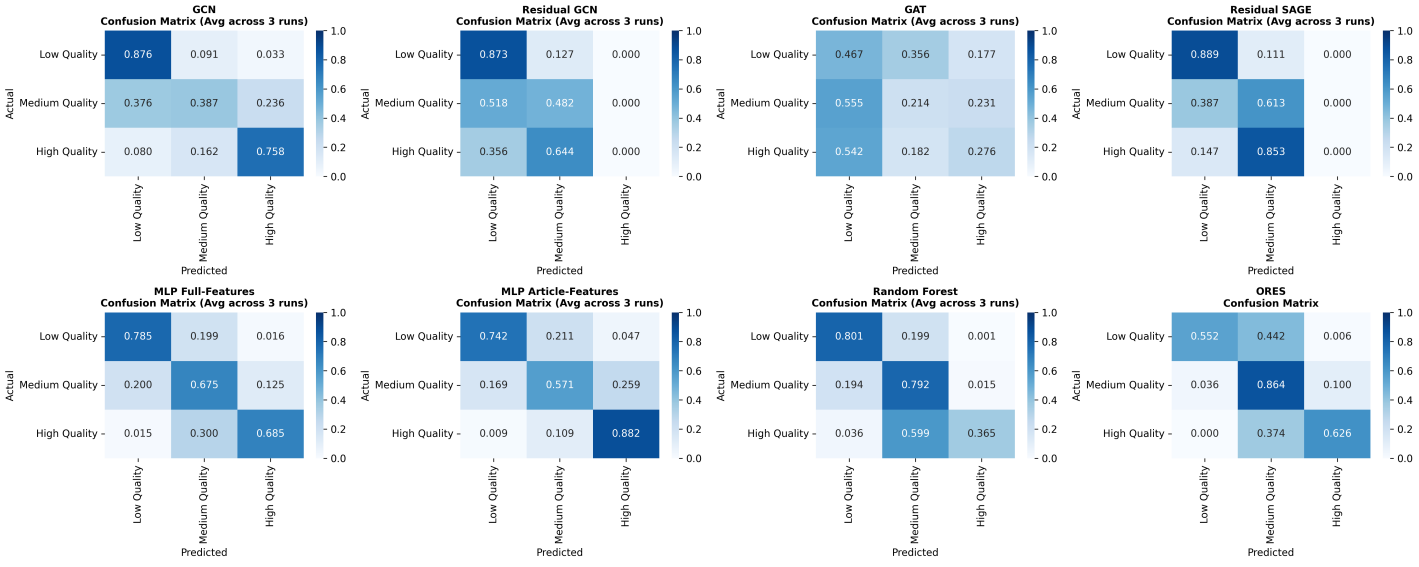



Figure 12: Confusion matrices for all models

By contrast, the neural models tend to favor the dominant Low Quality class. They perform reliably on Low

Quality articles but often misclassify High Quality ones, leading to weaker recall for this minority group. This is especially visible for GAT, which struggles overall, likely because Wikipedia's link network does not provide the kind of homophily its attention mechanism relies on.

ORES shows a different pattern: while it reaches decent accuracy (64.12%), its predictions are uneven across classes, with weaker performance on Low and High Quality. This suggests that ORES may be tuned toward other goals than our three-class setup.

#### *Which Features matter?*

The Random Forest analysis lets us gauge which features drive quality predictions. Content structure dominates - page_length (18.51%), num_references (12.67%) and num_sections (7.78%) are the top predictors.

Network features also contribute meaningfully: degree_out_centrality (3.40%), transition_max_prob (3.13%), and hub scores (2.45%) all rank in the top 10. The spectral embeddings capture additional structural patterns that explicit network metrics miss. Editorial activity like total edits (2.76%) provides another quality signal.

The permutation importance results largely agree with Gini importance for the top features, confirming page_length and num_references as robust predictors across different measurement approaches.

## *Discussion and Conclusion*

The comparison across models highlights both the potential and the limits of network-based approaches for Wikipedia quality prediction. Sampling only parts of the graph likely introduced bias, and the strong class imbalance, especially the very small number of high-quality articles made the task challenging for all models. While more extensive testing of architectures and parameters could refine results, the main issue is structural: Wikipedia's link network shows little homophily, so quality does not propagate directly through connections. This explains why GNNs provided no meaningful advantage over MLPs.

At the same time, the benchmarks reveal that network context is not irrelevant. The MLP with network features outperformed the article-only version by several percentage points, and the Random Forest feature importance analysis showed that measures like out-degree or hub scores, while secondary, still add value alongside dominant structural predictors such as page length and number of references. This reflects a broader distinction from academic citation networks: internal Wikipedia links are added mainly for navigation rather than endorsement. Consequently, link-based features alone cannot predict quality, but article position in the network can still provide useful complementary signals.

Overall, our findings suggest that while GNNs may not be the ideal model for this task under current conditions, combining structural and positional information remains a promising direction, particularly if future work addresses class imbalance and employs whole-network sampling strategies.

# References

## Code and Data

### API Endpoints

- https://wikimedia.org/api/rest_v1/metrics/pageviews/: Retrieved monthly pageview statistics
- https://en.wikipedia.org/api/rest_v1/: Accessed article summaries and last edit timestamps
- https://en.wikipedia.org/w/api.php: Batch-retrieved article metadata
- https://wikimedia.org/api/rest_v1/metrics/edits/: Collected detailed edit history data
- https://ores.wikimedia.org/: Benchmark

### Important Python Packages

- Data Processing & Machine Learning:
  - pandas & polars
  - scikit-learn
- Data manipulation and analysis:
  - numpy
  - umap-learn
- Graph Analysis & Neural Networks:
  - graph-tool
  - torch & torch-geometric
  - optuna
- Visualization & Results:
  - matplotlib & seaborn
  - tqdm
- Specialized Libraries:
  - scipy
  - concurrent.futures
  - requests

**Repository** The complete codebase, documentation, and reproducible experiments are available at: https://github.com/Ari-manius/DLSS-WAQP

***Literature***


Arroyo-Machado, Wenceslao, Daniel Torres-Salinas, Enrique Herrera-Viedma, and Esteban Romero-Frías. 2020. "Science Through Wikipedia: A Novel Representation of Open Knowledge Through Co-Citation Networks." *PloS One* 15 (2): e0228713.

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.

Cristian Consonni, David Laniado, and A. Montresor. 2019. "WikiLinkGraphs: A Complete, Longitudinal and Multi-Language Dataset of the Wikipedia Link Networks." *International Conference on Web and Social Media*. https://doi.org/10.1609/icwsm.v13i01.3257.

Han, Yi, Shanika Karunasekera, and Christopher Leckie. 2020. "Graph Neural Networks with Continual Learning for Fake News Detection from Social Media." arXiv. https://doi.org/10.48550/arXiv.2007.03316.

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.

Ruprechter, Thorsten, Tiago Santos, and Denis Helic. 2020. "Relating Wikipedia Article Quality to Edit Behavior and Link Structure." *Applied Network Science* 5 (1): 61.

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.