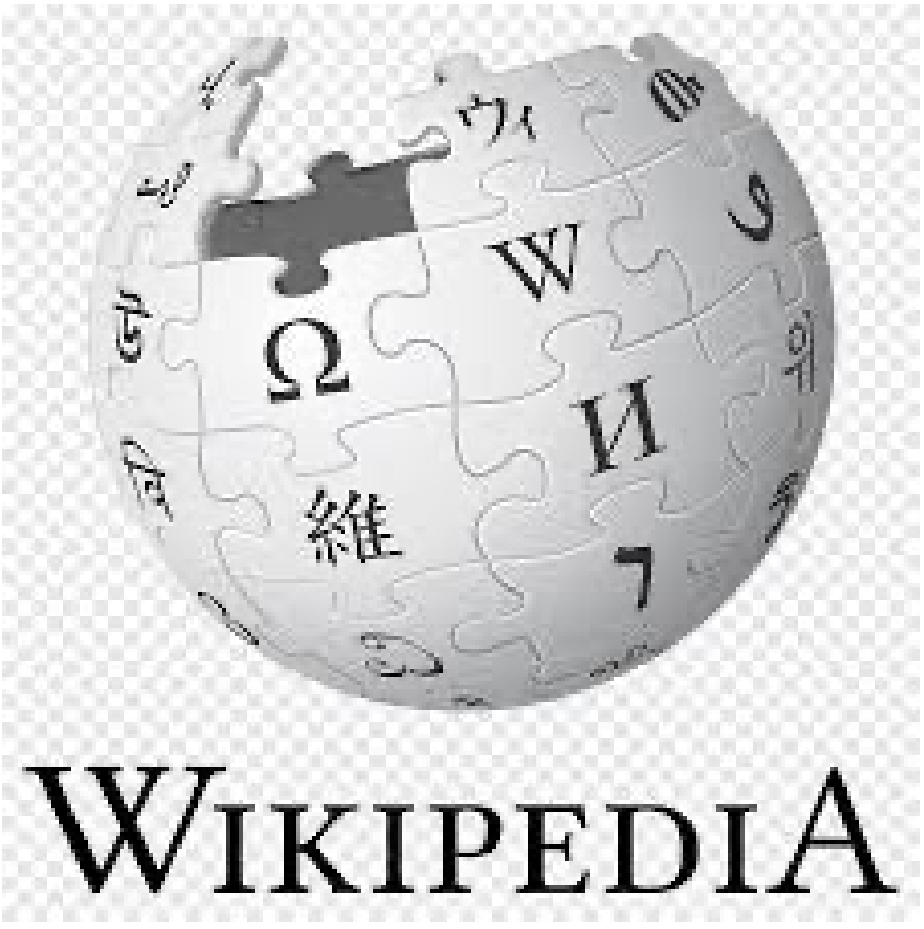
Wikipedia Article Quality Prediction

Final Report: DLSS

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

2025-08-11



Introduction

(Problem description and approach)

Citing this way: (2020), (2019)

Data Analysis

Dataset Description

(Features and Target Variable, Article based)

Wikipedia Network

Graph Description

Table 1: 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

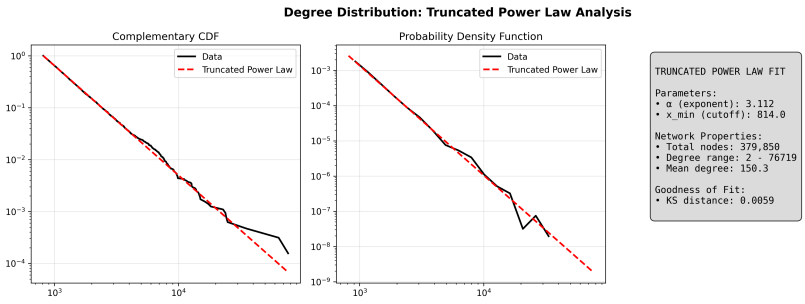


Figure 1: Wikipedia Graph - Degree Distribution and Power Law Analysis

Table 2: Network Assortativity

Assortativity	Value
Degree	0.0042
Quality Aggregated	0.0243
Quality	0.0139
Scalar Numeric Quality (Log)	0.0392
Scalar Numeric Quality	0.0312

Graph based Features

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

Methods

(Model architectures, training procedures, evaluation metrics)

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

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.

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.