Group 29
ExDDS WS24

Reproducing: “Topic Modeling on Podcast Short-Text Metadata” by Valero et al.

YAHYA JABARY*, TU Wien, Austria
JOACHIM BIBERGER, TU Wien, Austria
SELINA REINHARD, TU Wien, Austria
NUSAIBA AHMED, TU Wien, Austria
JULIA CHALISSERY, TU Wien, Austria

In this study, we set out to reproduce the evaluation of the NEiCE algorithm as detailed by Valero et al. in their paper "Topic Modeling on Podcast Short-Text Metadata". Our focus was on replicating the results using the Deezer and iTunes datasets. We found that our reproduction of the Deezer dataset scores was fairly successful, albeit with some minor discrepancies. However, the reproduction of the iTunes dataset scores revealed significant differences from the original results. These findings suggest that while the NEiCE algorithm’s performance on the Deezer dataset can be reliably reproduced, further investigation is needed to understand the variations observed in the iTunes dataset. Our study highlights the importance of reproducibility in research and the need for transparent reporting of methods and results to ensure the reliability of scientific findings.

ACM Reference Format:
Yahya Jabary, Joachim Biberger, Selina Reinhard, Nusaiba Ahmed, and Julia Chalissery. 2024. Reproducing: “Topic Modeling on Podcast Short-Text Metadata” by Valero et al. 1, 1 (December 2024), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Strategy

Score: $$CV(k) = \frac{1}{T}\sum_{i=1}^{T}\cos(v_{NPMI}(t_i), v_{NPMI}(t_{1:T}))$$

Paper hyperparams:

- @10 cutoff
- REL confidence: 0.9
- K: { 20, 50, 100, 200 }
- Alpha Word: { 0.2, 0.3, 0.4, 0.5 }
- Alpha Ent: { 0.3, 0.4 }

Assumed hyperparams:

- Number of topics: 10, 20, 50, 100 (common NMF arguments)
- Number of neighbors: 5, 10, 20, 500 (common CluWords arguments)

Encountered difficulties

Not resolved:

- Spotify dataset doesn’t exist
- Weights are not byte aligned
- Dependencies unknown
- Dependencies don’t work on ARM architecture
- GPU not supported (training took 3 full days)

Resolved:

- Reproducible Docker compose
- Seeds
- Lots of bug fixes in code

Key findings

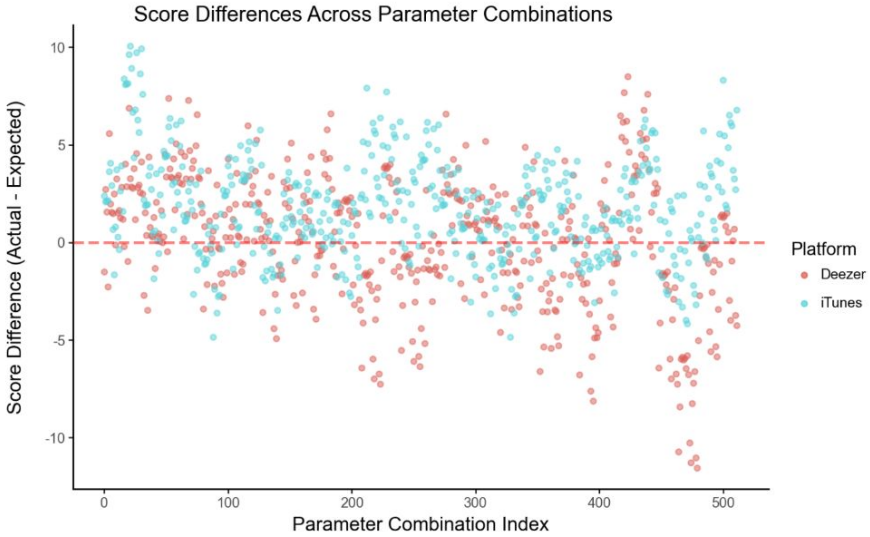
Deezer Scores:

T-test: statistic=0.228, p-value=0.819
Confidence intervals (95.0\%):
Actual: (51.994, 52.353)
Expected: (51.894, 52.383)
Variances:
Actual: 4.255
Expected: 7.909
KS-test: statistic=0.203, p-value=1.130e-09

iTunes Scores:

T-test: statistic=14.412, p-value=5.3177e-43
Confidence intervals (95.0\%):
Actual: (51.117, 51.406)
Expected: (49.345, 49.719)
Variances:
Actual: 2.751
Expected: 4.607
KS-test: statistic=0.452, p-value=2.0551e-47
iTunes Scores Statistics:
Actual scores count: 511
Non-null actual scores: 511
Sample of actual scores: [51.6682, 51.369, 51.4062, 51.432, 52.913]

Conclusion



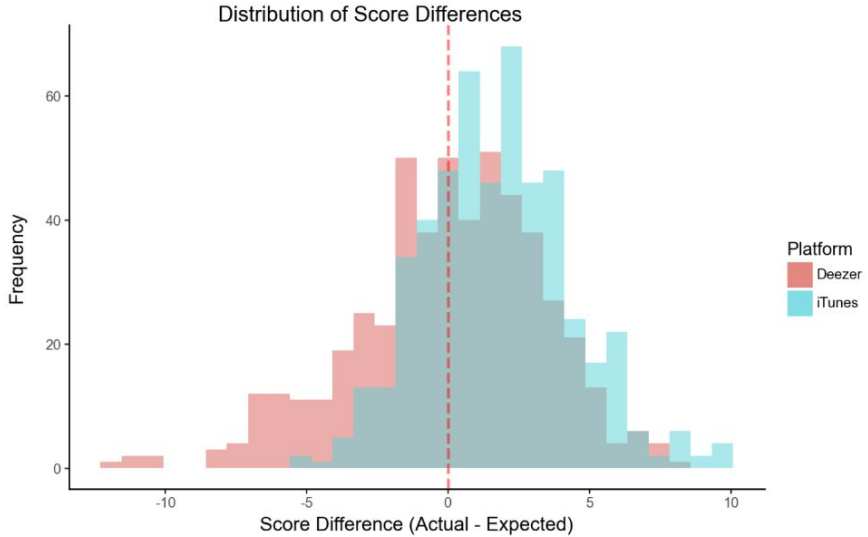


Fig. 1. Distribution of differences between our reproduced C_V scores and the authors' reported scores.

Conclusion

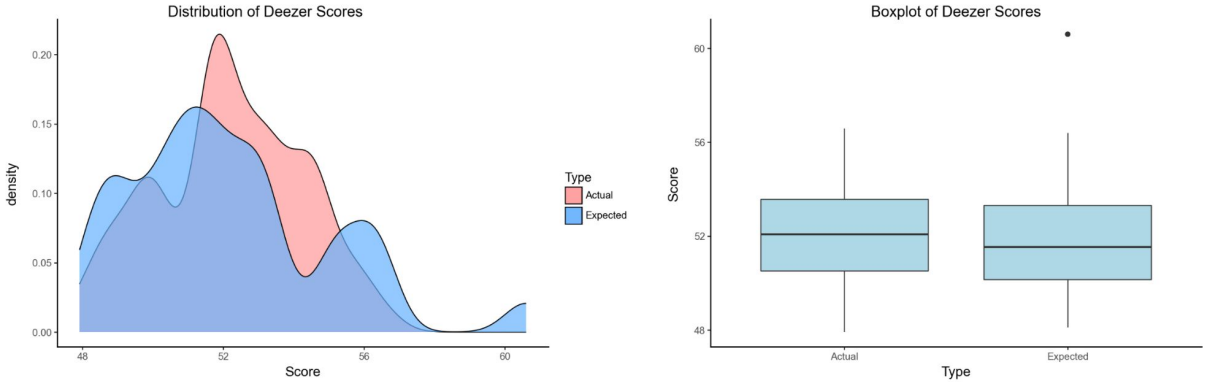


Fig. 2. Distribution of expected and actual C_V scores for the Deezer dataset.

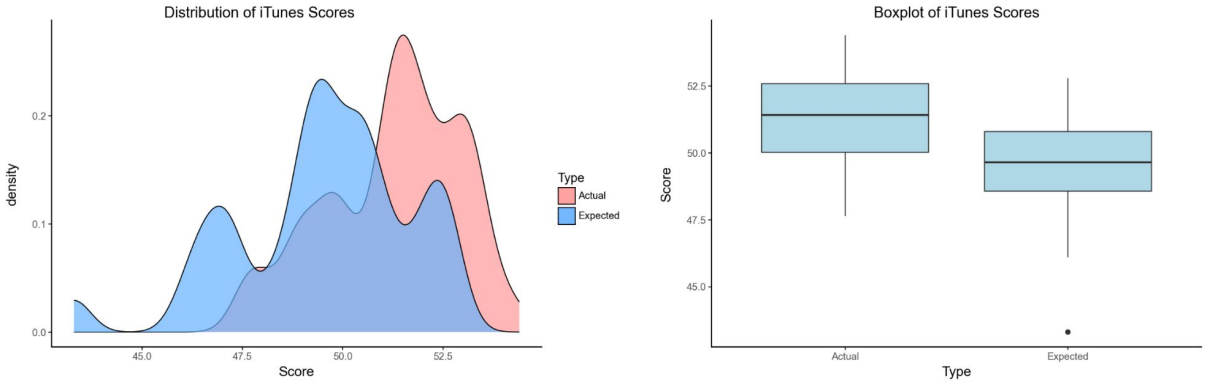


Fig. 3. Distribution of expected and actual C_V scores for the iTunes dataset.