Reproducing "Topic Modeling on Podcast Short-Text Metadata"

Authors: Anabel Dautović, Medak Mirta, Sharma Hrithik, Bosse Behrens, Felix Neuwirth

this is the abstract

**Additional Keywords and Phrases:** Podcast short-text metadata, …

1. Introduction

Podcasts have become a popular medium for sharing educational, entertaining, and informational content, with over 2 million podcasts and 48 million episodes available globally by 2021. Effective categorization is increasingly important but challenging due to the short, noisy, and sparse nature of podcast metadata and the unreliability of creator-assigned genre labels.

To address these challenges, Valero et al., in their ECIR 2022 study, "Topic Modeling on Podcast Short-Text Metadata" [1], introduced the Named-Entity-informed Corpus Embedding (NEiCE) model. NEiCE extends the CluWords approach by incorporating Named Entities (NEs) into a Non-negative Matrix Factorization (NMF) framework. This integration enables the model to leverage entity-related semantic information, improving topic coherence compared to state-of-the-art (SOTA) methods.

The original study evaluated NEiCE on three podcast datasets: Deezer, Spotify, and iTunes. The methodology included four key steps: extracting and linking named entities to Wikipedia entries using the Radboud Entity Linker (REL), preprocessing metadata to filter non-English entries with fastText and adapt JSON outputs for NEiCE, applying the NEiCE model to extend metadata with semantically related words, and evaluating topic coherence using the CV metric with Palmetto and Wikipedia as the external corpus.

This report reproduces the methodology and results of the original study to validate the effectiveness of NEiCE for improving topic coherence. By replicating the experimental workflow, we assess the feasibility of the model, identify reproducibility challenges—including issues with code, data access, and documentation—and explore its potential for broader applications in podcast metadata topic modeling.

1. Methodology

Our reproduction followed the following strategy. First, we thoroughly reviewed the original methodology, datasets, and code. We updated deprecated code, resolved dataset discrepancies, and automated parameter tuning to test different configurations of alpha\_word and alpha\_ent. Computational experiments were conducted using Docker containers to ensure consistent execution environments, with evaluations comparing coherence scores to the original study.

* 1. Data

The experiments were conducted using three distinct datasets, each comprising podcast metadata such as titles and descriptions in English. The Deezer dataset was created by the original authors and is publicly accessible. The iTunes dataset is also still accessible and included in our analysis. Lastly, the Spotify dataset, sourced from Spotify, is no longer available since the authors have stopped granting access requests.

* 1. Experiment setup

1. Results
   1. Results

Our results:

| **Topics** | **Dataset** | **(NEiCE,0.2, 0.3)** | **(NEiCE,0.2, 0.4)** | **(NEiCE,0.3, 0.3)** | **(NEiCE,0.3, 0.4)** | **(NEiCE,0.4, 0.3)** | **(NEiCE,0.4, 0.4)** | **(NEiCE,0.5, 0.3)** | **(NEiCE,0.5, 0.4)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20 | deezer | 52.130000 | 51.470000 | 51.240000 | 51.613333 | 54.090000 | 52.680000 | 53.050000 | 51.043333 |
| iTunes | 50.120000 | 51.793333 | 52.536667 | 52.430000 | 51.420000 | 52.353333 | 48.786667 | 50.523333 |
| 50 | deezer | 51.373333 | 52.866667 | 51.710667 | 52.133333 | 51.096667 | 52.700000 | 50.666667 | 52.500000 |
| iTunes | 50.373333 | 51.666667 | 51.656667 | 53.506667 | 52.556667 | 50.633333 | 50.866667 | 51.013333 |
| 100 | deezer | 51.886667 | 49.726667 | 50.746667 | 50.260000 | 54.580000 | 51.160000 | 49.973333 | 52.286667 |
| iTunes | 49.350000 | 52.080000 | 50.410000 | 50.196667 | 50.750000 | 50.163333 | 50.940000 | 50.980000 |
| 200 | deezer | 51.820000 | 48.156667 | 51.716667 | 49.500000 | 51.243333 | 51.043333 | 50.591667 | 51.926667 |
| iTunes | 49.340000 | 50.656667 | 51.616667 | 49.603333 | 50.250000 | 50.406667 | 48.840000 | 52.283333 |

Their results:

| **Topics** | **Dataset** | **(NEiCE,0.2, 0.3)** | **(NEiCE,0.2, 0.4)** | **(NEiCE,0.3, 0.3)** | **(NEiCE,0.3, 0.4)** | **(NEiCE,0.4, 0.3)** | **(NEiCE,0.4, 0.4)** | **(NEiCE,0.5, 0.3)** | **(NEiCE,0.5, 0.4)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20 | deezer | 50.2 | 53.1 | 48.5 | 53.3 | 53.2 | 56.4 | 52.5 | 56.3 |
| iTunes | 49.3 | 47.2 | 50.3 | 52.5 | 52.8 | 52.4 | 50.6 | 50.5 |
| 50 | deezer | 48.9 | 49.2 | 52.1 | 50.9 | 51.5 | 52.6 | 56.3 | 60.6 |
| iTunes | 43.3 | 49.5 | 52.5 | 49.5 | 50.1 | 51.9 | 46.5 | 52.0 |
| 100 | deezer | 51.4 | 50.8 | 51.5 | 55.3 | 52.2 | 48.1 | 50.8 | 54.9 |
| iTunes | 49.5 | 50.7 | 49.0 | 49.2 | 50.6 | 49.9 | 46.7 | 48.7 |
| 200 | deezer | 48.4 | 50.6 | 49.8 | 51.6 | 50.0 | 49.0 | 55.4 | 53.3 |
| iTunes | 47.0 | 51.3 | 48.2 | 49.8 | 51.1 | 47.4 | 49.0 | 46.1 |

Evaluation:

| **Topics** | **Dataset** | **(NEiCE,0.2, 0.3)** | **(NEiCE,0.2, 0.4)** | **(NEiCE,0.3, 0.3)** | **(NEiCE,0.3, 0.4)** | **(NEiCE,0.4, 0.3)** | **(NEiCE,0.4, 0.4)** | **(NEiCE,0.5, 0.3)** | **(NEiCE,0.5, 0.4)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20 | deezer | False | False | False | False | False | False | False | False |
| iTunes | False | False | False | False | False | False | False | False |
| 50 | deezer | False | False | False | False | False | False | False | False |
| iTunes | True | False | False | False | False | False | False | False |
| 100 | deezer | False | False | False | False | False | False | False | False |
| iTunes | False | True | False | False | False | False | False | False |
| 200 | deezer | False | False | False | False | False | False | False | False |
| iTunes | False | False | False | False | False | False | False | False |

* 1. Evaluation

1. Challenges and limitations

The reproduction process had some challenges and limitations. The main issue we faced was data accessibility, as the Spotify dataset isn’t available anymore, so the analysis was limited to Deezer and iTunes. The code had a few problems, such as outdated functions, missing dependencies like Gensim, and incomplete documentation, which required some debugging and updates. Pretrained entity linker models needed a lot of memory (over 60GB), and the process was computationally demanding, making experiments time-consuming. Some scripts for merging JSON outputs, adjusting preprocessing, and parameter tuning were missing, so we had to implement them ourselves. There were also some unclear preprocessing steps, minimal discussion of hyperparameter selection, and the evaluation was focused mainly on the CV metric, without scripts for a more thorough evaluation.

1. COnclusion

Reproducing the paper highlighted NEiCE's potential to improve topic coherence in short-text metadata while exposing critical challenges in reproducibility. Despite code and data limitations, our results aligned closely with the original study's trends, validating NEiCE’s effectiveness.

REFERENCES

1. Francisco B. Valero, Marion Baranes, and Elena V. Epure. 2022. Topic Modeling on Podcast Short-Text Metadata. In Advances in Information Retrieval, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.).