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

...

Additional Key Words and Phrases: Reproducibility, Information Retrieval

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), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

In this paper, we aim to reproduce the evaluation of the NEiCE algorithm as presented by Valero et al. [3].

1 Introduction

Named Entity informed Corpus Embedding (NEiCE), is a topic modeling algorithm that uses named entities (NE) to improve the quality of topics extracted from short-text content. The algorithm was introduced in the paper "Topic Modeling on Podcast Short-Text Metadata" by Valero et al. [3] from Deezer Research. It is based on the CluWords [4] algorithm, which clusters words based on their nearest neighbors. The authors claim that NeICE outperforms other topic modeling algorithms of the same class, such as Non-negative matrix factorization (NMF), Short-text topic modeling via non-negative matrix factorization (SeaNMF) [2] and Clustering words (CluWords) [4], on podcast metadata from Deezer, Spotify and iTunes.

The motivation behind this study stems from the nature of podcast metadata, which typically consists of short text such as titles and descriptions. Traditional topic modeling algorithms often struggle with short text due to the lack of context and sparse data, even when concatenated into pseudo-documents. However, NMF-based algorithms have shown promise in handling short text more effectively compared to probabilistic models, such as the Generalized Polya Urna Dirichlet Multinomial Mixture (GPU-DMM) [1]. Additionally, NMF-based algorithms offer better interpretability than neural models, such as the Negative sampling and Quantization Topic Model (NQTM) [5]. NEiCE offers an improvement over CluWords by leveraging named entities, which are more informative and coherent than regular words.

Authors' Contact Information: Yahya Jabary, jabaryyahya@gmail.com, TU Wien, Austria; Joachim Biberger, TU Wien, Austria; Selina Reinhard, TU Wien, Austria; Nusaiba Ahmed, TU Wien, Austria; Julia Chalissery, TU Wien, Austria.

Manuscript submitted to ACM

1.1 NEiCE Algorithm

The NEiCE algorithm consists of a (1) preprocessing stage and a (2) topic modeling stage. The preprocessing stage prepares the podcast metadata for topic modeling by enriching the text with named entities, generating contextualized word embeddings and creating document-level representations. The topic modeling stage then applies Non-negative Matrix Factorization (NMF) to discover relevant topics from the preprocessed data.

1.1.1 Preprocessing Stage. The preprocessing stage begins with named entity recognition (NER) in both the title and description fields of podcast metadata. The system employs the Radboud Entity Linker (REL) library for linking these entities to Wikipedia entries. Within this linking process, the Flair library, which utilizes embeddings, performs Named Entity Recognition (NER) to detect entity mentions. It then uses wikipedia2vec to identify unique candidates from a list of potential matches.

For each identified named entity, the system generates a Wikipedia page reference along with a confidence score. This score is crucial as it determines how the entity will be treated in subsequent processing – either as a simple text span, as a confirmed named entity or as individual words requiring separate processing.

The vocabulary cleaning phase incorporates the NameDataset library to eliminate overly common named entities such as actors and athletes that might not contribute meaningfully to topic identification. Following the cleaning process, the system applies a specialized NE-related re-weighting to the term frequency (tf) factor.

The preprocessing stage handles different datasets with specific criteria. For both iTunes and Spotify datasets, the system removes duplicate titles and eliminates entries where the combined title and description contain fewer than three terms. The Spotify dataset undergoes additional language filtering, utilizing both fastText and CLD3 to ensure only English-language podcasts are retained. The Deezer dataset, which is the largest among the three, contains similar metadata including creator-provided information, titles, descriptions and show names in English. Across all datasets, genres with fewer than 300 shows are excluded to ensure robust topic modeling.

1.1.2 Topic Modeling Stage. The Topic Modeling Stage uses Non-negative Matrix Factorization (NMF) as the primary technique. It begins with the creation of a document-term matrix X, where each row represents a podcast and each column represents a term from the vocabulary. The matrix X is then factorized into two non-negative matrices W and H, such that $X \approx WH$. In this factorization, W represents the document-topic matrix, while H represents the topic-term matrix.

To enhance the topic modeling process, the authors propose the novel Entity-informed Contextual Embedding (NEiCE). This representation leverages named entities, which are often present in podcast metadata, to improve the quality of discovered topics. It combines contextual word embeddings with named entity information, allowing for a more nuanced and accurate topic modeling process.

The optimization problem for NMF is formulated as minimizing the Frobenius norm of the difference between X and WH, subject to non-negativity constraints on W and H. This is expressed mathematically as $\min(W, H \geq 0) \cdot \|X - WH\|^2 F$. The resulting factorization provides a low-rank approximation of the original document-term matrix, with each column of W representing a topic and each row of H representing the importance of terms within that topic.

1.2 Evaluation

The evaluation primarily uses CV topic coherence, which correlates best with human judgment of topic ranking. The CV score is calculated as:

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

where T is the number of top words per topic, NPMI vectors are computed using Wikipedia as the external corpus and the final coherence is averaged across all K topics.

The hyperparameters specified in the original paper are as follows:

- Number of top words (T): { 10 }
- Number of topics (K): { 20, 50, 100, 200 }
- REL confidence threshold: 0.9 for named entity linking
- α^{word} range: { 0.2, 0.3, 0.4, 0.5 }
- α^{ent} range: { 0.3, 0.4 }

The experiments were conducted on an Intel Xeon Gold 6134 CPU @ 3.20GHz with 32 cores and 128GB RAM.

The table 1 shows the topic coherence scores C_V obtained by NEiCE for each (α^{word},α^{ent}) configuration on the Deezer, Spotify and iTunes datasets.

We want to reproduce this table and conduct a statistical analysis to evaluate our confidence in the results.

Dataset	Deezer				Spotify				iTunes			
	20	50	100	200	20	50	100	200	20	50	100	200
NEiCE (0.2, 0.3)	50.2	48.9	51.4	48.4	51.7	49.0	45.2	46.5	49.3	43.3	49.5	47.0
NEiCE (0.2, 0.4)	53.1	49.2	50.8	50.6	48.7	48.7	43.5	41.7	47.2	49.5	50.7	51.3
NEiCE (0.3, 0.3)	48.5	52.1	51.5	49.8	52.2	49.0	47.5	47.6	50.3	52.5	49.0	48.2
NEiCE (0.3, 0.4)	53.3	50.9	55.3	51.6	50.1	48.5	51.1	49.8	52.5	49.5	49.2	49.8
NEiCE (0.4, 0.3)	53.2	51.5	52.2	50.0	53.2	49.5	50.5	45.9	52.8	50.1	50.6	51.1
NEiCE (0.4, 0.4)	56.4	52.6	48.1	49.0	51.0	48.2	47.3	47.8	52.4	51.9	49.9	47.4
NEiCE (0.5, 0.3)	52.5	56.3	50.8	55.4	51.3	47.7	45.6	45.4	50.6	46.5	46.7	49.0
NEiCE (0.5, 0.4)	56.3	60.6	54.9	53.3	55.0	49.9	46.7	45.0	50.5	52.0	48.7	46.1

Table 1. Topic coherence scores C_V (in %) obtained by NEiCE for each (α^{word},α^{ent}) configuration on the Deezer, Spotify and iTunes datasets – named “Table 5” in the original paper.

2 Strategies

3 Difficulties

4 Key Findings

5 Conclusion

References

[1] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR ’16)*. Association for Computing Machinery, New York, NY, USA, 165–174. https://doi.org/10.1145/2911451.2911499

[2] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 world wide web conference*. 1105–1114.

[3] Francisco B Valero, Marion Baranes, and Elena V Epure. 2022. Topic modeling on podcast short-text metadata. In *European Conference on Information Retrieval*. Springer, 472–486.

[4] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 753–761. https://doi.org/10.1145/3289600.3291032

[5] Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1772–1782.

A System Specifications

All experiments were conducted on a consumer-grade laptop with the following specifications:

```
$ system_profiler SPSoftwareDataType SPHardwareDataType
Software:

    System Software Overview:

      System Version: macOS 14.6.1 (23G93)
      Kernel Version: Darwin 23.6.0
      Boot Volume: Macintosh HD
      Boot Mode: Normal
      Computer Name: Yahya's MacBook Pro
      User Name: Yahya Jabary (sueszli)
      Secure Virtual Memory: Enabled
      System Integrity Protection: Enabled
      Time since boot: 103 days, 2 hours, 59 minutes

Hardware:

    Hardware Overview:

      Model Name: MacBook Pro
      Model Identifier: Mac14,10
      Model Number: <redacted>
      Chip: Apple M2 Pro
      Total Number of Cores: 12 (8 performance and 4 efficiency)
      Memory: 16 GB
      System Firmware Version: 10151.140.19
      OS Loader Version: <redacted>
      Serial Number (system): <redacted>
      Hardware UUID: <redacted>
      Provisioning UDID: <redacted>
      Activation Lock Status: Disabled
```