



Petrov, A. and Macdonald, C. (2022) A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation. In: ACM Conference on Recommender Systems (RecSys 2022), Seattle, USA, 18-23 Sep 2022, pp. 436-447. ISBN 9781450392785 (doi: [10.1145/3523227.3548487](https://doi.org/10.1145/3523227.3548487))

Publisher's URL: <https://dl.acm.org/doi/10.1145/3523227.3548487>

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

Copyright © 2022 Association for Computing Machinery. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. ACM Conference on Recommender Systems (RecSys 2022), Seattle, USA, 18-23 Sep 2022, pp. 436-447. ISBN 9781450392785

<http://eprints.gla.ac.uk/275645/>

Deposited on: 28 July 2022

Enlighten – Research publications by members of the University of
Glasgow

<http://eprints.gla.ac.uk>

A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation

Aleksandr Petrov
a.petrov.1@research.gla.ac.uk
University of Glasgow
United Kingdom

Craig Macdonald
Craig.Macdonald@glasgow.ac.uk
University of Glasgow
United Kingdom

ABSTRACT

BERT4Rec is an effective model for sequential recommendation based on the Transformer architecture. In the original publication, BERT4Rec claimed superiority over other available sequential recommendation approaches (e.g. SASRec), and it is now frequently being used as a state-of-the-art baseline for sequential recommendation. However, not all subsequent publications confirmed its superiority and have proposed other models that were shown to outperform BERT4Rec in effectiveness. In this paper we systematically review all publications that compare BERT4Rec with another popular Transformer-based model, namely SASRec, and show that BERT4Rec results are not consistent within these publications. To understand the reasons behind this inconsistency, we analyse the available implementations of BERT4Rec and show that we fail to reproduce results of the original BERT4Rec publication when using their default configuration parameters. However, we are able to replicate the reported results with the original code if training for a much longer amount of time (up to 30x) compared to the default configuration. We also propose our own implementation of BERT4Rec based on the HuggingFace Transformers library, which we demonstrate replicates the originally reported results on 3 out of 4 datasets, while requiring up to 95% less training time to converge. Overall, from our systematic review and detailed experiments, we conclude that BERT4Rec does indeed exhibit state-of-the-art effectiveness for sequential recommendation, but only when trained for a sufficient amount of time. Additionally, we show that our implementation can further benefit from adapting other Transformer architectures that are available in the HuggingFace Transformers library (e.g. using disentangled attention, as provided by DeBERTa, or larger hidden layer size cf. ALBERT). For example, on the MovieLens-1M dataset, we demonstrate that both these models can improve BERT4Rec performance by up to 9%. Moreover, we show that an ALBERT-based BERT4Rec model achieves better performance on that dataset than state-of-the-art results reported in the most recent publications.

ACM Reference Format:

Aleksandr Petrov and Craig Macdonald. 2022. A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22)*, September 18–23, 2022, Seattle, WA, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3523227.3548487>

RecSys '22, September 18–23, 2022, Seattle, WA, USA

© 2022 Association for Computing Machinery.

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Sixteenth ACM Conference on Recommender Systems (RecSys '22)*, September 18–23, 2022, Seattle, WA, USA, <https://doi.org/10.1145/3523227.3548487>.

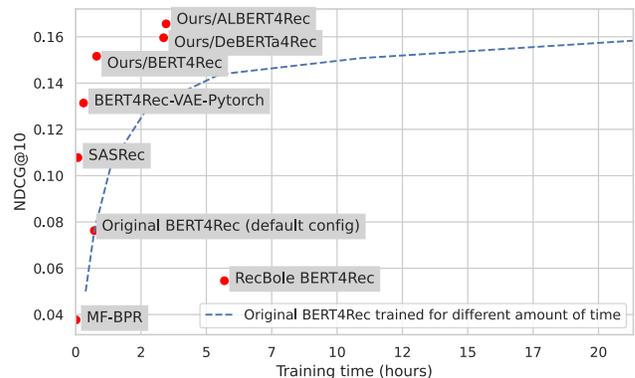


Figure 1: Comparison of four implementations of BERT4Rec (Original, RecBole, BERT4Rec-VAE-Pytorch, Ours/BERT4Rec) with two baseline models (MF-BPR, SASRec) and two more advanced Transformer-based models (Ours/ALBERT4Rec, Ours/DeBERTa4Rec) on the MovieLens-1M dataset. All points represent models with their default configuration. The dashed line represents NDCG@10 of the original BERT4Rec implementation trained for different amounts of training time.

1 INTRODUCTION

Sequential Recommender Systems are a class of recommender systems that use the order of interactions to make recommendations. The goal of such recommendation models is to predict the next user-item interaction by the sequence of user's past interactions. This goal is also known as *next item prediction task*. While early sequential recommender systems applied Markov Chains [50], more recently neural networks based models have been shown to outperform traditional models [20, 39, 54, 61]. Since the arrival of the Transformer neural architecture [57] and, in particular the BERT [11] language model, Transformer-based sequential recommendations models, such as SASRec [24], S³Rec [69], LightSANS [12], NOVA-BERT [35] and DuoRec [47] have achieved state-of-the-art performance in next item prediction.

In particular, BERT4Rec is a highly-cited Transformer-based sequence recommendation model proposed by Sun et al. [52], which adapts the BERT language model, by equating tokens with items and sentences with sequences of item interactions. Sun et al. [52] argued that, one of the main advantages of BERT4Rec compared to other Transformer-based models, such as SASRec, is that it uses an item masking training task [11] – this argument was further investigated

and confirmed by Petrov and Macdonald [44]. Applied to recommendation, the main idea of this training task is to replace random items in the training sequences with a special *[mask]* token and train the model to recover these masked tokens; we describe details of this task in Section 3.3. In the original publication [52], BERT4Rec was claimed to achieve significant superiority over existing neural and traditional approaches, however subsequent publications by a number of different authors (e.g. [13, 62, 69]) did not confirm this superiority. Our goal therefore is to understand the reasons behind this discrepancy. In particular, we analyse three open source implementations of BERT4Rec with an aim to reproduce the originally reported results using these implementations. We find that there is a marked discrepancy in both effectiveness and efficiency of these implementations. We also demonstrate that training the original implementation of BERT4Rec, with its default configuration, results in an underfitted model, and it requires up to 30x more training time in order to replicate the originally reported results.

Ultimately, we show that, with proper configuration, BERT4Rec achieves better performance than earlier models such as SASRec, and that some results reported in subsequent papers are based on underfitted versions of BERT4Rec. Moreover, we show that an appropriately trained BERT4Rec can match or outperform later models (e.g. DuoRec [47], LightSANs [12] & NOVA-BERT [35]) and therefore may still be used as a state-of-the-art sequential recommendation model.

In addition, we propose our own implementation of BERT4Rec that is based on a popular HuggingFace Transformers library [59]. HuggingFace Transformers is a popular machine learning library with a large supporting community. For instance, its GitHub repository¹ has almost 62K stars and 15K forks, which makes it the second most popular machine learning library on GitHub.² It has become a de facto standard for publishing Transformer-based models and therefore contains well optimised and efficient versions of Transformer-based models. Based on this and the fact that it was already successfully applied for recommendations [10], we expect that using it as a backbone for building Transformer-based recommenders should be both effective and efficient. Indeed, we show that our implementation with default configuration replicates the results published by Sun et al. [52] and requires up to 95% less time to achieve these results. Moreover, the BERT architecture in our implementation can be easily replaced with other models available in the HuggingFace Transformers library. We demonstrate by using two examples (DeBERTa [15] and ALBERT [27]) that such replacement can lead to further improved effectiveness.

Figure 1 summarises the results of our findings on the MovieLens-1M dataset. Each point on the figure corresponds to a default configuration of a recommendation model and the dashed line represents performance of the original BERT4Rec code when trained for different amounts of time. The figure shows that, depending on the chosen BERT4Rec implementation, the observed NDCG@10 effectiveness can vary from 0.0546 for RecBole implementation to 0.156 (our implementation, 3x difference). It also shows that the required training time also depends on the implementation. For example, the BERT4Rec-VAE implementation requires 18 minutes to train with default configuration whereas the original implementation requires

44 minutes while achieving poorer NDCG@10 than BERT4Rec-VAE. The figure also show that the original implementation can reach the same level of performance as the best implementation, however it requires much more training compared to what is specified in the default configuration. The figure also portrays results of our implementation based on DeBERTa [15] and ALBERT [27] Transformers and shows that these models can outperform BERT4Rec, however they require more time than our BERT4Rec implementation. Nevertheless, their training times are still 85% smaller than what is required for the full convergence of the original BERT4Rec implementation.

In short, the contributions of the paper are: (1) a systematic review of the papers comparing BERT4Rec and SASRec, which shows that the results of such comparisons are not consistent; (2) an analysis of the available implementations of BERT4Rec, which shows that frequently these implementations fail to reproduce results reported in [52] when trained with their default parameters, (3) an analysis of impact of the training time to recommendation performance of the original BERT4Rec implementation; (4) a new HuggingFace-based implementation of BERT4Rec that successfully replicates the results reported in [52]; (5) two new recommendation models: DeBERTa4Rec and ALBERT4Rec that improve quality of our implementation via replacing BERT model with DeBERTa and ALBERT respectively. The code and the documentation for this paper can be found in the virtual appendix in our GitHub repository.³

The rest of the paper is organised as follows: Section 2 formally introduces the sequential recommendation problem and the main approaches to solve it; Section 3 describes details of SASRec and BERT4Rec, popular Transformer-based sequential recommendations models; Section 4 contains a systematic review of papers comparing SASRec and BERT4Rec; Section 5 describes available BERT4Rec implementations; Sections 6 and 7 describe experimental setup and the results of the experiments respectively; Section 8 put this work in context, wrt. related work and performances observed in recent publications; Section 9 summarises the results and contains final remarks.

2 BACKGROUND

The goal of the sequential recommendation is to predict the next interaction in a sequence. Consider a set of users $u \in U$, where each user has made a sequence of interactions $s \in S = \{i_1, i_2, i_3 \dots i_t\}$, and $i_k \in I$ denotes items. The task of the recommender system is to predict next element i_{t+1} in the sequence. More formally, the goal of the sequential recommender system M is to produce a list of K items ranked by the probability of being the sequence continuation:

$$M(s) \rightarrow \{i_{s_1}, i_{s_2}, i_{s_3}, i_{s_4} \dots i_{s_K}\}$$

Some of the first sequential models were based on Markov Chains [50], however over the last few years the best models for sequential recommendations are based on neural networks. Early neural network-based architectures for sequential recommendations included recurrent neural networks [19, 20] and convolutional neural networks [54, 61], however, following the advancements in natural language processing, more recent models have adopted the

¹ <https://github.com/huggingface/transformers> ² After Tensorflow, according to <https://github.com/EvanLi/Github-Ranking>

³ https://github.com/asash/bert4rec_repro

self-attention mechanism [30, 36] and the Transformer architecture [24, 45, 52, 69]. Transformer-based models consistently outperform both traditional (non-neural) models, as well as earlier neural architectures [24, 52, 69].

Two popular Transformer-based sequential recommendation models are SASRec [24] and BERT4rec [52]. The main difference between these two models is the training task: during training, SASRec only uses past interactions to predict future ones, whereas BERT4Rec adopts items masking, which allows to use both past and future interactions to recover masked items. We describe the details of these models and their training tasks in Section 3. BERT4Rec was published one year later than SASRec, and in the original publication it uses SASRec as a baseline. In the original publication, the authors reported improvements over SASRec on all four datasets they used for comparison. Indeed, for example by the NDCG@10 metric, the authors reported 14.02% improvement on the Beauty dataset, 5.31% on the Steam dataset, 10.32% on the MovieLens-1M dataset, and 14.47% on the MovieLens-20M dataset. However, these results were not consistently confirmed in some subsequent publications: for example Zhou et al. [69] reported -5.73% NDCG@10 decrease on Beauty dataset, Qiu et al. [47] reported -47.01% on MovieLens-1M and Yue et al. [62] reported -0.8% on Steam. This paper analyses the reasons for this discrepancy. In the next section, we describe the architectures of the SASRec and BERT4rec models.

3 SASREC AND BERT4REC

3.1 Transformer

Both SASRec and BERT4Rec are based on the Transformer [57] architecture, which was, originally designed for natural language processing. The original Transformer architecture consists of the encoder and decoder parts, however recommendation models use only the encoder part of a Transformer.

Essentially, a Transformer encoder is a sequence-to-sequence model, which translates a sequence of tokens to a sequence of vectors. It encodes the sequence of the input tokens using an embedding layer, sums these embeddings with positional embeddings (vectors encoding position of the token in the sequence), and passes them through several Transformer *blocks*. A Transformer block is the core component of the architecture that consists of a multi-head self attention layer, a pointwise feed-forward layer and residual connections. We refer to the original publication [57] for more details about the architecture. In order to use this architecture, SASRec and BERT4Rec use sequences of item ids as the inputs. Each input sequence represents a history of interactions of a single user. An output of the Transformer encoder is a sequence of vectors. Both SASRec and BERT4Rec project each vector from the encoder output sequence to the score distribution over items via multiplying these vectors by the matrix of the learnable item embeddings. Figure 2 illustrates architectures of both models. We now describe the details of the two models and highlight their key differences.

3.2 SASRec

Figure 2a illustrates the SASRec model. The model uses a sequence of items (a history of interactions of a single user) as its input and predicts the same sequence shifted by one element to the right.

This means, that the last element in the predicted sequence corresponds to the future next iteration, which exactly matches next item prediction - the main task of the sequential recommendation. Indeed, at the inference time, SASRec only uses the last element in the predicted sequence.

SASRec uses the causal (unidirectional) version of self attention: to predict element i_k of the output sequence it can only access elements $1, 2, \dots, i_k$ from the input sequence. Casual self attention prevents the model from experiencing information leakage: with regular (bidirectional) self attention, the task would be trivial - the model would just copy element i_{k+1} from the input sequence as its k^{th} output.

3.3 BERT4Rec

Figure 2b illustrates BERT4Rec architecture. As we can see, it is very similar to the SASRec architecture. One important difference is that BERT4Rec uses regular (bidirectional) self attention instead of casual attention. BERT4Rec also uses a different training task - instead of predicting the shifted sequence, it employs an item masking training task. Specifically, at the training time, some of the items in the input sequences are replaced with a special *[mask]* token. The goal of the model is to recover these masked tokens. Items masking task allows to generate more training samples - up to $\binom{n}{k}$, where n is the sequence length and k is the number of masked items - out of a single sequence. At the inference time, BERT4Rec adds a *[mask]* token to the end of the sequence of interactions and then for each item in the catalogue BERT4Rec computes probability of being this *[mask]* token.

Items masking is loosely connected to the next item prediction task, which makes it harder to train. However, as shown in the original BERT4Rec publication [52], its bidirectional nature may be advantageous and improve the quality of the model. Nevertheless, as we show in the next section, the question whether or not BERT4Rec is a superior model compared to SASRec is still an open question and recent publications do not give a definitive answer to this question.

4 SYSTEMATIC REVIEW OF SASREC AND BERT4REC PERFORMANCE

The results published in the original BERT4Rec paper [52] are indeed promising. According to the publication, BERT4Rec outperformed the second best model, SASRec, on four different datasets (Beauty [16], Steam [24], MovieLens-1M [14] and MovieLens-20M [14]) and across six different evaluation metrics (Recall@{1, 5, 10}, NDCG@{5, 10} and MRR). In order to validate this result and estimate its reproducibility, we perform a systematic review of the papers that compare SASRec and BERT4Rec. We formulate two hypotheses that we test in our systematic review:

- H1** Overall, BERT4Rec is not systematically better than SASRec in the published literature.
- H2** Even when experimental setup is similar (e.g. experiment on the same datasets), the outcome of the comparison of BERT4Rec and SASRec may be different, which indicates poor replicability of BERT4Rec results.

Hypothesis H1 addresses a general question if BERT4Rec is indeed a superior model when compared to SASRec; i.e. it outperforms

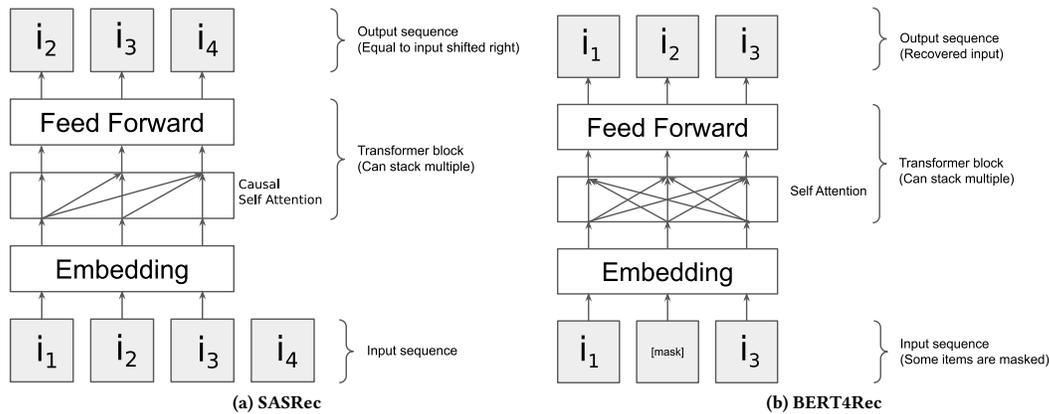


Figure 2: Architectures of SASRec & BERT4Rec. SASRec uses casual (unidirectional) Self Attention, whereas BERT4rec uses regular (bidirectional) Self Attention. SASRec aims to predict the input sequence shifted right, whereas BERT4Rec recovers masked items.

SASRec across a large number of different datasets and metrics. Hypothesis H2 addresses the question of reproducibility; we look in particular datasets and metrics and check if the experiment results are consistent within this experimental setup.

4.1 Review Methodology

To test Hypotheses H1 and H2 we conducted a review of all papers citing the original BERT4Rec publication [52] according to the Google Scholar website.⁴ From these papers, we chose to include only those papers that used both BERT4Rec and SASRec models for their experiments, and examined their observed performances. We further excluded papers that were not peer reviewed to minimise chances of relying on non-verified experiments.⁵ We considered performing a *meta analysis*, by aggregating improvements across multiple papers (e.g. “average improvement by BERT4Rec in NDCG@10 over SASRec”). However, we observe that the identified papers applied various evaluation measures (e.g. varying rank cutoffs) and methodology (e.g. different negatives sampling strategies, see Section 6.4). Instead we considered three outcomes of the experiments: “BERT4Rec wins”, “SASRec wins” and “Tie”, and relied on counting. In particular, we counted that a particular model wins a comparison if it was better according to all metrics used in the experiment. If a model was better according to one subset of metrics and worse according to another, we counted a tie.

To test the overall superiority of BERT4Rec over SASRec (Hypothesis H1), we determine the total numbers of each possible outcome (BERT4Rec wins; SASRec wins; Tie) and check whether or not there is any considerable amount of situations when SASRec wins over BERT4Rec. To test results reproducibility on different datasets (Hypothesis H2), we aggregate the results by datasets and count experiment outcomes for each dataset independently.

4.2 Systematic Review Results

We reviewed the 370 papers citing BERT4Rec according to Google Scholar on 25/03/2022.⁶ Following the inclusion criteria (comparison with SASRec) we found 58 publications containing such comparisons, and thereafter excluded 18 (not peer-reviewed etc.). This left a total of 40 publications that compare BERT4Rec and SASRec, making a total with 134 total comparisons on 46 different datasets (3.35 comparisons per paper on average, 2.91 comparisons per dataset on average).

4.2.1 H1. Overall results consistency. Table 1 summarises the results of our systematic review. As we can see from the “Total” row of the table, BERT4Rec indeed wins the majority of the comparisons (86 out of 134, 64%). However, the numbers of comparisons, which SASRec won (32 out of 134, 24%) or at least achieved a tie (16 out of 134, 12%) is not negligible: 36% overall. Therefore we can conclude that BERT4Rec is not consistently superior compared to SASRec in the published literature, which validates Hypothesis H1 and is not consistent with the original BERT4Rec paper. We now investigate if this inconsistency can be explained by the differences in datasets, i.e. if the reason is that BERT4Rec is better on some datasets, but SASRec is better on others.

4.2.2 H2. Poor replicability of BERT4Rec. We turn again to Table 1 and analyse the results of the systematic review aggregated by dataset. The table includes all datasets from our review appearing in at least 5 papers. From the table, we observe that the proportion of outcomes is indeed highly dependent on the dataset. For example, on the both ML-20M and Steam datasets, BERT4Rec won 7 out 8 comparisons (88%), whereas on Sport it won just 1 out of 6 (17%) and on Toys it won 0 out of 5 (0%) comparisons. Therefore, it appears that the disparity in the overall results can be explained by some salient characteristics of the datasets.

However there is still some level of inconsistencies within datasets. For example, on the two most popular datasets, Beauty and ML-1M, the proportion of experiments which BERT4Rec did not

⁴ <http://scholar.google.com> ⁵ Other examples of excluded publications include a paper withdrawn by authors – and only available in the web archive – and an MSC student thesis.

⁶ The spreadsheet of analysed papers is also included in our virtual appendix.

win roughly matches the overall result: 7 out of 19 for Beauty (37%) and 5 out of 13 for ML-1M (28%). Importantly, the original BERT4Rec paper [52] found that BERT4Rec was superior by a statistically significant margin on both of these datasets, however we can see a large number of papers failed to replicate this result, which validates our Hypothesis H2.

Overall, we argue that the lack of agreement in the observed results is problematic: it means that papers have used BERT4Rec with different configurations and ultimately many of these papers may be using poor configurations of BERT4Rec as their baselines. This leads us to investigate the available open source implementations of BERT4Rec and their hyperparameter settings.

5 BERT4REC IMPLEMENTATIONS

To understand the reasons for the poor replicability of BERT4Rec results, demonstrated by our systematic review, we analyse the available BERT4Rec implementations. During our systematic review, we identified three available BERT4Rec implementations cited in the papers: (1) the original implementation, provided by the authors of the original paper [52]; (2) RecBole [67] - a library that contains a large number of recommendation models, including BERT4Rec (3) BERT4Rec-VAE - a project that implements BERT4Rec and Variational Autoencoder [34] models using PyTorch.⁷

In addition to these three existing implementations, we note the Transformers4Rec project [10], which is based on the popular HuggingFace Transformers natural language processing library [59]. Transformers4Rec does not include an implementation of BERT4Rec, however, inspired by this project we implement our own version of BERT4Rec that uses HuggingFace’s version of BERT as a backbone. Indeed, given the popularity of the HuggingFace Transformers library, we expect that models implemented using it will be highly efficient and may outperform other available implementations.

Salient characteristics of all four BERT4Rec implementations are listed in Table 2. We now turn to the experiments with these four implementations, which help us to understand poor replicability of BERT4Rec results.

6 EXPERIMENTAL SETUP

6.1 Research Questions

We aim to address following research questions with our experiments:

- RQ1** Can we replicate the state-of-the-art results reported in [52] using all of the available BERT4Rec implementations, applying their default configurations?
- RQ2** What is the effect of the training time on the performance of the original BERT4Rec implementation?
- RQ3** Can our HuggingFace-based implementation of BERT4Rec benefit from replacing BERT with another language model available in the HuggingFace Transformers library?

⁷ <https://github.com/jaywonchung/BERT4Rec-VAE-Pytorch>

6.2 Datasets

Because our goal is to examine replicability of the BERT4Rec, we use same four datasets as used in the original publication [52], namely ML-1M [14], Beauty [16], Steam [24] and ML-20M [14]. These datasets also represent 4 of the 5 most popular datasets in our systematic review. To ensure quality of the data and following the common practice [17, 24, 50, 54], Sun et al. [52] discarded sequences with less than 5 interactions. In order to stay consistent with their experimental setup, for ML-1M, Steam, and Beauty we use the preprocessed versions of the datasets from the original BERT4Rec repository.⁸ Table 3 presents salient characteristics of our experimental datasets.

We split each dataset into train, validation and test partitions using a Leave-One-Out strategy: for each user, we hold the final interaction for the test set; for a validation set, we select the second last action for 2048 randomly selected users from each dataset; the remaining interactions are used for training.

6.3 Models

For our replicability experiment (RQ1), we use four BERT4Rec implementations, as described in Section 5. For all four implementations, we apply their default data preprocessing pipelines and default model parameters. Table 4 lists the default parameter values. The original BERT4Rec version includes slightly different parameters tuned for each different datasets and we use the ones specified for the ML-20M dataset; however according to the original publication [52] these changes in parameters within the limits specified in the configuration files have only a limited effect and therefore should not change overall model performance. For the ML-1M and ML-20M datasets we also experiment with a so-called “longer seq” version, where we increase maximum sequence length, because the average number of items per user is much larger in these two datasets. Overall, the configuration of our implementation of BERT4Rec is similar to the original model, with a notable difference in the training stopping criteria: to ensure that the models are fully trained, we use an early stopping mechanism; we measure the value of loss function on the validation data and stop training if validation loss did not improve for 200 epochs.

We also use two baselines in our replicability experiments: (1) MF-BPR [49] - a classic matrix factorisation based approach with a pairwise BPR loss. We use the implementation of this model from the LightFM library [26] and set the number of latent components to 128; (2) SASRec [24] - a Transformer-based model, described in Section 3.2. We use an adaptation of the original code for this model.⁹ For SASRec we set sequence length to 50, embedding size to 50 and use 2 transformer blocks; according to the experiments conducted by Kang et al. [24], these parameters are within the range where SASRec shows reasonable performance.

In the extra training experiment (RQ2) we use the default configuration of the original BERT4Rec implementation with the exception of number of training steps, which we vary between 200,000 and 12,800,000.

Finally, to compare BERT with other models available in the HuggingFace Transformers library (RQ3), we experiment with two recent Transformer-based architectures: (1) DeBERTa [15] - a model

⁸ <https://github.com/FeiSun/BERT4Rec/tree/master/data>

⁹ <https://github.com/kang205/SASRec>

Table 1: Results of BERT4Rec vs SASRec comparisons in the peer-reviewed publications. Bold denotes model with more wins on a dataset. Asterisk (*) denotes datasets used in the original BERT4Rec paper [52]. Only the datasets with appearing in at least 5 papers are presented (8 datasets out of 46), however the comparison results on all other 38 datasets are included in the "Total" numbers.

dataset	total	BERT4Rec wins	SASRec wins	Ties	BERT4Rec wins papers	SASRec wins papers	Ties papers
Beauty* [16]	19	12 (63%)	5 (26%)	2 (11%)	[52], [40], [68], [38], [29], [7], [64], [55], [62], [18], [60], [53]	[69], [32], [65], [58], [47]	[3], [9]
ML-1M* [14]	18	13 (72%)	3 (17%)	2 (11%)	[52], [40], [68], [29], [7], [23], [55], [18], [60], [51], [42], [46], [28]	[12], [62], [47]	[64], [9]
Yelp [2]	10	6 (60%)	4 (40%)	0 (0%)	[69], [1], [3], [58], [47], [42]	[12], [32], [65], [33]	
Steam* [43]	8	7 (88%)	1 (12%)	0 (0%)	[52], [40], [68], [29], [64], [9], [60]	[62]	
ML-20M* [14]	8	7 (88%)	0 (0%)	1 (12%)	[52], [40], [66], [29], [7], [60], [46]		[9]
Sports [16]	6	1 (17%)	4 (67%)	1 (17%)	[28]	[69], [32], [65], [47]	[3]
LastFM [5]	6	4 (67%)	2 (33%)	0 (0%)	[69], [23], [55], [28]	[1], [65]	
Toys [16]	5	0 (0%)	5 (100%)	0 (0%)		[69], [13], [32], [65], [3]	
Total	134	86 (64%)	32 (23 %)	16 (12 %)			

Table 2: BERT4Rec implementations used in our experiments.

Implementation	GitHub URL	Framework	GitHub stars	Example papers
Original	FeiSun/BERT4Rec	Tensorflow v1	390	[21, 22, 37]
RecBole	RUCAIBox/RecBole	PyTorch	1,800	[3, 12, 31]
BERT4Rec-VAE	jaywonchung/BERT4Rec-VAE-Pytorch	PyTorch	183	[48, 56, 63]
Ours/HuggingFace	asash/bert4rec_repro	Tensorflow v2	N/A	N/A

Table 3: Experimental datasets.

Dataset	Users	Items	Interactions	Avg. len.	Sparsity
ML-1M	6,040	3,416	999,611	165.49	0.9515
Steam	281,428	13,044	3,488,885	12.398	0.9990
Beauty	40,226	54,542	353,962	8.79	0.9998
ML-20M	138,493	26,744	20,000,263	144.41	0.9946

that improves BERT with a disentangled attention mechanism [15] where each word is encoded using two vectors (a vector for content and a vector for position); (2) ALBERT [27] - a model that improves BERT via separating the size of the hidden state of the vocabulary embedding from the number of the hidden layers. It also introduces cross-layer parameter sharing, which allows to decrease overall

number of parameters in the network. Following the BERT4Rec naming convention, we call these two models DeBERTa4Rec and ALBERT4Rec. For these two models we use same model configuration parameters as for BERT4Rec with the exception of longer sequence length, which for which we considered values from {50, 100, 200} and report the results for the best-performing value (200) according to NDCG@10 metric on validation set.

6.4 Metrics

Following Sun et al. [52], we evaluate the models using the Leave-One-Out strategy and focus on two ranking-based metrics: NDCG and Recall@K on the test data.¹⁰ The authors report results using

¹⁰ The original BERT4Rec publication uses the name Hit Rate (HR) instead of Recall. Hit Rate represent probability of successful prediction of one single interaction in the next-item prediction task. In case of sequential recommendation, when there is only one true positive item per user, this metric is equivalent to Recall, and we prefer the more conventional name.

sampled metrics: for each positive item they sample 100 negative items - these 101 items are then ranked for evaluation. In particular, they use a popularity-based sampling of negatives: the probability of sampling an item as a negative is proportional to its overall popularity in the dataset. However, sampled metrics are known to be problematic. Indeed, it has been shown in a number of recent publications [4, 9, 25] that sampled metrics are not always consistent with full, unsampled versions (where all items are ranked for each user) and can lead to incorrect model comparisons. Nevertheless, because our goal is to examine the reproducibility of the BERT4Rec model, we apply sampled metrics in to compare our results with the results reported in the original paper. In addition, following the recommendations in [25] and [4], we report results on *unsampled metrics*, where we rank all items from the dataset at evaluation time. However, in experiments where we do not need to compare with the results reported in the original publication [52] we only report unsampled metrics.

6.5 Criteria for a Successful Replication

There is some level of randomness involved in deep neural networks training. Indeed, even when the same model is trained multiple times, random weights initialisation and random data shuffling can lead to small changes in the resulting effectiveness metrics. BERT4Rec also randomly masks items during training, which also slightly increases the randomness of the process. Furthermore, as mentioned above, the evaluation metrics used by the original BERT4Rec publication [52] involve sampling random items and therefore may change even when evaluating exactly same model.

Thus, it is almost impossible to exactly replicate the reported results. Therefore, we need to define some interval around the values reported in [52] within which we can say that our model replicates originally reported results. Because we only know a single measurement per dataset, we can not rely on hypothesis testing methods, such as paired t-test, so we need a heuristic for defining the tolerance interval.

According to Madhyastha and Jain [41], the standard deviation in classification metrics for instances of the same deep neural model trained with different random seeds can reach 2.65%. If we assume that ranking metrics, such as NDCG and Recall, exhibit similar variance under different model instances, this suggests that a larger tolerance is needed for defining equivalence. Hence, in this paper we define that a *model replicates the results reported in the original publication on a metric, if the metric value is equal to the originally reported value within a relative tolerance of $\pm 5\%$.*

7 EXPERIMENTAL RESULTS

7.1 RQ1. Replicability with default configurations.

We first analyse if the available implementations of BERT4Rec are able to replicate the originally reported results, when trained with their default configurations. Table 5 compares the results achieved by the models trained with their default configuration with the results reported in the original BERT4Rec paper [52]. As we can see from the table, both versions of our HuggingFace-based implementation replicate the reported results on 3 datasets out of 4 (all except Beauty) for the popularity-sampled Recall@10 metric and

on 2 datasets out of 4 for the popularity-sampled NDCG@10 metric. Moreover, BERT4Rec-VAE replicates the originally reported results on 2 datasets for the popularity-sampled Recall@10 metrics (both versions of MovieLens) and on one dataset for NDCG@10 metric (ML-1M). However, the original and RecBole implementations of BERT4Rec fail to replicate the originally reported results on all four datasets by a large margin (e.g. original implementation is 46.47% worse according to Recall@10 metric on Steam).

The table also reports comparison of the BERT4Rec implementations with the two baseline recommender models. As we can see, the best version of BERT4Rec always statistically significantly (according to the 2-tailed t-test with Bonferroni multiple test correction, $pvalue < 0.05$) outperforms both MF-BPR and SASRec on all four datasets and on both sampled and unsampled metrics. BERT4Rec-VAE performs better than the baselines on all four datasets, and the Ours and Ours (longer seq) versions of BERT4Rec perform better than baselines on three out of four datasets (except Beauty). SASRec performs better than the original and RecBole implementations on three datasets out of four datasets (except Beauty), which echoes the inconsistencies identified in our systematic review. Further, comparisons of the RecBole and original implementations with MF-BPR on Steam and ML-20M can lead to different conclusions if we look to sampled or unsampled metrics. For example according to sampled Recall@10, the RecBole implementation is 33% worse than MF-BPR on Steam dataset, however at the same time it is 86% better than MF-BPR according to full Recall@10. This discrepancy is in line with the results reported in [4, 9, 25] and shows importance of using unsampled metrics in the research.

From the table, we also see that there is big discrepancy in training times.¹¹ For example, on the ML-1M dataset, training BERT4Rec-VAE takes only 18 minutes, whereas training the RecBole version takes 5.6 hours. At the same time, BERT4Rec-VAE achieves 2.3 times better performance than RecBole, according to unsampled Recall@10.

Unfortunately, we fail to replicate results reported in [52] for the Beauty dataset. In attempts to achieve originally reported results, we also ran experiments with the original implementation with a configuration tuned for this specific dataset, tried training of the original model for up to 16x more training time, and tried using the evaluation framework and metrics from the original codebase. None of these attempts allowed us to reach metric values significantly better than we obtained from the BERT4Rec-VAE implementation, which are 22.68% worse than the reported values.

Overall, the answer to the RQ1 depends on the implementation and on the dataset. According to popularity-sampled Recall@10, we can replicate the originally reported results with our implementation in 3/4 cases, in 2/4 cases using BERT4Rec-VAE and in 0/4 cases using original and RecBole implementations. Furthermore, for unsampled metrics, we can observe similar conclusions: BERT4Rec models that fail to replicate the originally reported (sampled) results also performed poorly on unsampled metrics as well.

¹¹ We report training times on our hardware configuration: 16 (out of 32) cores of an AMD Ryzen 3975WX CPU; NVIDIA A6000 GPU; 128GB memory.

Table 4: Default parameters of the BERT4Rec implementations.

Implementation	Original	RecBole	BERT4Rec-VAE	Ours	Ours (longer seq)
Sequence length	200	50	100	50	100
Training stopping criteria	400,000 steps	300 epochs	200 epochs	Early stopping: 200 epochs	Early stopping: 200 epochs
Item masking probability	0.2	0.2	0.15	0.2	0.2
Embedding size	64	64	256	64	64
Transformer blocks	2	2	2	2	2
Attention heads	2	2	4	2	2

7.2 RQ2 Effects of training time on the original BERT4Rec implementation performance

Sun et al. [52] did not report the amount of training they needed to reach the state-of-the-art metrics. The amount of training specified in the original BERT4Rec code is controlled by the parameter *training steps*, which is set to 400,000 by default. To understand the effect of training time, we vary the number of training steps from slightly less than default (200,000, 0.5x compared to the default) to much larger (1,280,000, 32x compared to the default), which results in a variation of training time from 23 minutes to 21.3 hours. Figure 3 portrays the relationship between training time and test performance of the original BERT4Rec implementation on the ML-1M dataset using popularity-sampled metrics.¹² As can be observed from the figure, it is possible to reproduce the originally reported results using the original BERT4Rec implementation, however it requires many times more training time than the default configuration. Indeed, the model needs to be trained almost 11 hours instead of 42 minutes to replicate results on sampled Recall@10 (15x times) and 21 hours to replicate results on sampled NDCG@10 (30x times). Looking back at Figure 1, the effect of training time on unsampled version of NDCG follows the same general trend: the performance that the model reaches with the default configuration is 51.8% lower compared to what can be reach with 30x increased amount of training.

Overall in answer to RQ2, we find that the number of training steps set in the default configuration of the original BERT4Rec code is too small, and to reach performance levels reported by Sun et al. [52], the training time has to be increased up 30x times (from less than an hour to almost a day of training). This change makes the model much harder to train, and for example, makes hyperparameters tuning much less feasible, specifically with limited amount of hardware. It also makes it very easy to mistakenly use an underfitted version of BERT4Rec in the experiments. This likely explains inconsistencies observed in our systematic review.

7.3 RQ3 Other Transformers from HuggingFace for Sequential Recommendation

To answer our final research question, we compare our BERT4Rec implementation with two other models: DeBERTa4Rec (based on the DeBERTa [27] architecture) and ALBERT4Rec (based on the ALBERT [15] architecture). Both architectures propose improvements to the BERT architecture (some of the improvements we describe in

¹² We use popularity-sampled metrics, because we need to compare results with the values reported by Sun et al. [52]

Section 6.3), and both models were shown to outperform BERT on the GLUE benchmark in their native natural language processing domain. Our goal is to experiment if these improvements hold in the domain of sequential recommendations. Our HuggingFace based implementation makes such an experiment very simple, as in the Transformers library these models have compatible interfaces with BERT. We also apply the same early stopping mechanism as we use in our version of BERT4Rec (see Section 6.3).

The results for these new models on the ML-1M dataset are listed in Table 6. From the table, we observe that the both models are better outperform BERT4Rec, however only for ALBERT4Rec are the results statistically significant (+6.48% Recall@10, +9.23% NDCG@10). These improvements allow us to conclude that the answer to RQ3 is positive - our implementation of BERT4Rec can further other models available in HuggingFace Transformers library. Overall, there are more than 100 model architectures available in the Transformers library, and many of these models can be used seamlessly instead of BERT with our BERT4Rec implementations.

8 RELATED WORK & DISCUSSION

As observed in Section 4, there has been some difficulty in replicating the BERT4Rec reported results in [52]. In terms of related work, we highlight Dacrema et al. [8] who showed that deep learning-based recommendation papers frequently used weak configurations of baselines. They demonstrated this on an example of simpler matrix factorisation baselines, however in this work we show, that this problem also exists with more complicated (and effective) models. Chin et al. [6] showed that the dataset selection can also influence the conclusion about models performance, which echoes the results of our systematic review. Krichene et al. [25] showed that conclusions about model performances may change when sampled or unsampled metrics are used – this was recently specifically confirmed for sequential recommendation [9]. Our experiments also confirm these results: for example, on the MovieLens-1M dataset, a Matrix Factorisation model performs 12% better than the RecBole implementation of BERT4Rec according to the sampled version of Recall@10, but at the same time it is 30% worse according to the unsampled version of Recall@10. However, none of the aforementioned works performed a systematic review about one popular technique, nor examined the reasons why that technique could be difficult to replicate. Nevertheless, the answers to our research questions hold in both sampled and unsampled versions of the metrics.

Table 5: Replicability of the originally reported BERT4Rec results. Bold denotes a successful replication of a metric reported in [52], percentages show difference with the reported metric, underlined denotes the best model by a metric, † denotes a statistically significant difference with the best model on the paired t-test with Bonferoni multiple testing correction ($pvalue < 0.05$).

(a) ML-1M Dataset						
	Model	Popularity-sampled		Unsampled		Training Time
		Recall@10	NDCG@10	Recall@10	NDCG@10	
Baselines	MF-BPR	0.5134 (-26.34%)†	0.2736 (-43.21%)†	0.0740†	0.0377†	58
	SASRec	0.6370 (-8.61%)†	0.4033 (-16.29%)†	0.1993†	0.1078†	316
BERT4Rec versions	Original	0.5215 (-25.18%)†	0.3042 (-36.86%)†	0.1518†	0.0806†	2,665
	RecBole	0.4562 (-34.55%)†	0.2589† (-46.26%)†	0.1061†	0.0546†	20,499
	BERT4Rec-VAE	0.6698 (-3.90%)†	0.4533 (-5.29%)†	0.2394†	0.1314†	1,085
	Ours	0.6865 (-1.51%)	0.4602 (-4.48%)	0.2584	0.1392	3,679
	Ours (longer seq)	0.6975 (+0.07%)	0.4751 (-1.39%)	<u>0.2821</u>	<u>0.1516</u>	2,889
Reported [52]	BERT4Rec	0.6970	0.4818	N/A	N/A	N/A

(b) Steam Dataset						
	Model	Popularity-sampled		Unsampled		Training Time
		Recall@10	NDCG@10	Recall@10	NDCG@10	
Baselines	MF-BPR	0.3466 (-13.63%)†	0.1842 (-18.53%)†	0.0398†	0.0207†	162
	SASRec	0.3744 (-6.70%)†	0.2052 (-9.24%)†	0.1198†	0.0482†	3,614
BERT4Rec versions	Original	0.2148 (-46.47%)†	0.1064 (-52.94%)†	0.0737†	0.0375†	4,847
	RecBole	0.2325 (-42.06%)†	0.1177 (-47.94%)†	0.0744†	0.0377†	83,816
	BERT4Rec-VAE	0.3520 (-12.29%)†	0.1941 (-14.15%)†	0.1237†	0.0526†	65,303
	Ours	0.3978 (-0.87%)	0.2219 (-1.86%)	<u>0.1361</u>	<u>0.0734</u>	117,651
Reported [52]	BERT4Rec	0.4013	0.2261	N/A	N/A	N/A

(c) Beauty Dataset						
	Model	Popularity-sampled		Unsampled		Training Time
		Recall@10	NDCG@10	Recall@10	NDCG@10	
Baselines	MF-BPR	0.2090 (-30.91%)†	0.1089 (-41.47%)†	0.0185†	0.0090†	58
	SASRec	0.1111 (-63.27%)†	0.0524 (-71.83%)†	0.0079†	0.0036†	316
BERT4Rec versions	Original	0.1099 (-63.67%)†	0.0567 (-69.55%)†	0.0163†	0.0079†	3,249
	RecBole	0.1996 (-34.02%)†	0.1103 (-40.76%)†	0.0158†	0.0079†	11,024
	BERT4Rec-VAE	<u>0.2339 (-22.68%)</u>	<u>0.1407 (-24.44%)</u>	<u>0.0331</u>	<u>0.0188</u>	21,426
	Ours	0.1891 (-37.49)†	0.0919 (-50.64%)†	0.0166†	0.0080†	14,497
Reported [52]	BERT4Rec	0.3025	0.1862	N/A	N/A	N/A

(d) ML-20M Dataset						
	Model	Popularity-sampled		Unsampled		Training Time
		Recall@10	NDCG@10	Recall@10	NDCG@10	
Baselines	MF-BPR	0.6126 (-18.02%)†	0.3424 (-35.88%)†	0.0807†	0.0407†	197
	SASRec	0.6582 (-11.92%)†	0.4002 (-25.06%)†	0.1439†	0.0724†	3635
BERT4Rec versions	Original	0.4027 (-46.11%)†	0.2193† (-58.93%)†	0.0939†	0.0474†	6,029
	RecBole	0.4611 (-38.30%)†	0.2589 (-51.52%)†	0.0906†	0.0753†	519,666
	BERT4Rec-VAE	0.7409 (-0.86%)	0.5259 (-1.52%)	<u>0.2886</u>	<u>0.1732</u>	23,030
	Ours	0.7127 (-4.63%)†	0.4805 (-10.02%)†	0.2393†	0.1310†	44,610
	Ours (longer seq)	0.7268 (-2.74)†	0.4980 (-6.74%)†	0.2514†	0.1456†	39,632
Reported [52]	BERT4Rec	0.7473	0.5340	N/A	N/A	N/A

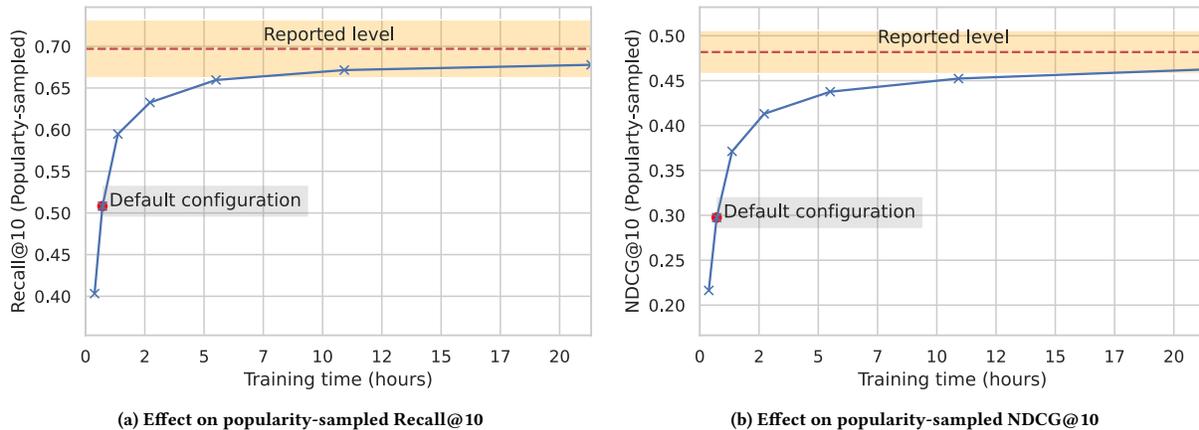


Figure 3: Effect of the training time on the performance of the original implementation of BERT4Rec. The dashed line represents the metric value reported in [52], filled area represents $\pm 5\%$ interval around the reported value, within which we count the result as a "replication" of the originally reported result. Default configuration corresponds to the results with the parameters specified in the original repository for the ML-1M dataset.

Table 6: Comparison of the Transformer models from the HuggingFace library. All results are reported on full (unsampled) metrics. The percentage shows the relative difference with BERT4Rec. Bold denotes the best model by a metric, \dagger denotes statistically significant difference compared with BERT4Rec according to a paired t-test with Bonferroni multiple testing correction ($pvalue < 0.05$).

Model	Recall@10	NDCG@10	Training Time
BERT4Rec	0.282	0.151	2,889
DeBERTa4Rec	0.290 (+3.0%)	0.159 (+2.3%)	12,114
ALBERT4Rec	0.300 (+6.4%)\dagger	0.165 (+9.2%)\dagger	12,453

Our results for RQ1 and RQ2 in Section 5 above suggest that available BERT4Rec implementations are the cause of this difficulty, probably due to underfitting. Indeed, some of the recent papers that seemingly have used an underfitted BERT4Rec as a baseline for evaluation on the ML-1M dataset are listed in Table 7. All these papers used unsampled metrics for evaluation, so we can compare their reported values with a fully-trained BERT4Rec results. As we can see from the table, the difference with our fully fitted version of BERT4Rec ranges from -10% to -53%, the results that are inline with the ones we observe when trained BERT4Rec implementations with default configurations. Moreover, the results reported for the best models in these papers are rather worse than those we observe for a fully-trained BERT4Rec (GRU4Rec+ -20.59%, LightSANs -19.03%) or only marginally better (NOVA-BERT +1.55%, DuoRec +4.43%) models. Furthermore, as shown Section 7.3, similar or even better improvements can be achieved by a simple replacement of one Transformer with another. According to this table, ALBERT4Rec improves the state-of-the-art result (+2% Recall@10 compared to DuoRec), however this improvement needs to be verified via direct

Table 7: BERT4Rec results and best model results reported in the literature for the ML-1M dataset. The results are copied from the respective publications. All results are reported on full (unsampled) metrics. The percentage shows the difference with our implementation of BERT4Rec. Bold denotes the best reported BERT4Rec result and the best result overall.

Publication	BERT4Rec Recall@10	Best model	Best model Recall@10
This paper	0.282 (+0.0%)	ALBERT4Rec	0.300 (+6.4%)
Dallmann et al. [9]	0.160 (-43.2%)	GRU4Rec+	0.224 (-20.5%)
Qiu et al. [47]	0.132 (-53.1%)	DuoRec	0.294 (+4.4%)
Fan et al. [12]	0.221 (-21.6%)	LightSANs	0.228 (-19.0%)
Liu et al. [35]	0.252 (-10.5%)	NOVA-BERT	0.286 (+1.5%)

comparison with statistical significance testing, rather than comparing reported numbers. We leave these experiments as well as experiments on more datasets and with other Transformer architectures available in the HuggingFace Transformers Library for future research.

9 CONCLUSIONS

In this paper we conducted a systematic review of the papers comparing BERT4Rec and SASRec and found that the reported results were not consistent among these papers. To understand the reasons of the inconsistency, we analysed the available BERT4Rec implementations and found that in many cases they fail to replicate originally reported results when trained with their default parameters. Furthermore, we showed that the original implementation requires much more training time compared to the default configuration in order to replicate originally reported results. This gives weight to the argument that in some cases papers have used underfitted versions of BERT4Rec as baselines.

We also proposed our own implementation of BERT4Rec based on the HuggingFace Transformers library, which in most of the cases replicates the originally reported results with the default configuration parameters. We showed that our implementation achieves similar results to the most recent sequential recommendation models, such as NOVA-BERT and DuoRec. We also showed our implementation can be further improved by using other architectures available in the HuggingFace Transformers library. We believe that this paper, as well as our openly available code, will help the researchers to use appropriately trained baselines and will move the science in the right direction.

REFERENCES

- [1] Mehrnaz Amjadi, Seyed Danial Mohseni Taheri, and Theja Tulabandhula. 2021. Katrec: Knowledge aware attentive sequential recommendations. In *Proc. ICDS*. 305–320.
- [2] Nabihah Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362* (2016).
- [3] Shuqing Bian, Wayne Xin Zhao, Kun Zhou, Jing Cai, Yancheng He, Cunxiang Yin, and Ji-Rong Wen. 2021. Contrastive Curriculum Learning for Sequential User Behavior Modeling via Data Augmentation. In *Proc. CIKM*. 3737–3746.
- [4] Rocío Cañameres and Pablo Castells. 2020. On target item sampling in offline recommender system evaluation. In *Proc. RecSys*. 259–268.
- [5] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2011. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In *Proc. RecSys* (Chicago, IL, USA) (RecSys 2011). New York, NY, USA.
- [6] Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The Datasets Dilemma: How Much Do We Really Know About Recommendation Datasets?. In *Proc. WSDM*. 141–149.
- [7] Sung Min Cho, Eunhyeok Park, and Sungjoo Yoo. 2020. MEANTIME: Mixture of attention mechanisms with multi-temporal embeddings for sequential recommendation. In *Proc. RecSys*. 515–520.
- [8] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proc. RecSys*. 101–109.
- [9] Alexander Dallmann, Daniel Zoller, and Andreas Hotho. 2021. A Case Study on Sampling Strategies for Evaluating Neural Sequential Item Recommendation Models. In *Proc. RecSys*. 505–514.
- [10] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4Rec: Bridging the Gap between NLP and Sequential/Session-Based Recommendation. In *Proc. RecSys*. 143–153.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*. 4171–4186.
- [12] Xinyan Fan, Zheng Liu, Jianxun Lian, Wayne Xin Zhao, Xing Xie, and Ji-Rong Wen. 2021. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proc. SIGIR*. 1733–1737.
- [13] Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S Yu. 2021. Continuous-time sequential recommendation with temporal graph collaborative transformer. In *Proc. CIKM*. 433–442.
- [14] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TOIS)* 5, 4 (2015), 1–19.
- [15] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. In *Proc. ICLR*.
- [16] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proc. WWW*. 507–517.
- [17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proc. WWW*. 173–182.
- [18] Zhankui He, Handong Zhao, Zhe Lin, Zhaowen Wang, Ajinkya Kale, and Julian McAuley. 2021. Locker: Locally Constrained Self-Attentive Sequential Recommendation. In *Proc. CIKM*. 3088–3092.
- [19] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proc. CIKM*. 843–852.
- [20] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *Proc. ICLR*.
- [21] Kaixi Hu, Lin Li, Qing Xie, Jianquan Liu, and Xiaohui Tao. 2021. What is Next when Sequential Prediction Meets Implicitly Hard Interaction?. In *Proc. CIKM*. 710–719.
- [22] Liwei Huang, Yutao Ma, Yanbo Liu, Bohong Danny Du, Shuliang Wang, and Deyi Li. 2021. Position-enhanced and Time-aware Graph Convolutional Network for Sequential Recommendations. *ACM Transactions on Information Systems (TOIS)* (2021).
- [23] Taegwan Kang, Hwanhee Lee, Byeongjin Choe, and Kyomin Jung. 2021. Entangled bidirectional encoder to autoregressive decoder for sequential recommendation. In *Proc. SIGIR*. 1657–1661.
- [24] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *Proc. ICDM*. 197–206.
- [25] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proc. KDD*. 1748–1757.
- [26] Maciej Kula. 2015. Metadata Embeddings for User and Item Cold-start Recommendations. In *Proc. Workshop on New Trends on Content-Based Recommender @ RecSys (CEUR Workshop Proc., Vol. 1448)*. 14–21.
- [27] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proc. ICLR*.
- [28] Chenglin Li, Mingjun Zhao, Huanming Zhang, Chenyun Yu, Lei Cheng, Guoqing Shu, BeiBei Kong, and Di Niu. 2022. RecGURU: Adversarial Learning of Generalized User Representations for Cross-Domain Recommendation. (2022), 571–581.
- [29] Haoyang Li, Xin Wang, Ziwei Zhang, Jianxin Ma, Peng Cui, and Wenwu Zhu. 2021. Intention-aware sequential recommendation with structured intent transition. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2021).
- [30] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proc. CIKM*. 1419–1428.
- [31] Yicong Li, Hongxu Chen, Xiangguo Sun, Zhenchao Sun, Lin Li, Lizhen Cui, Philip S Yu, and Guandong Xu. 2021. Hyperbolic hypergraphs for sequential recommendation. In *Proc. CIKM*. 988–997.
- [32] Yang Li, Tong Chen, Peng-Fei Zhang, and Hongzhi Yin. 2021. Lightweight self-attentive sequential recommendation. In *Proc. CIKM*. 967–977.
- [33] Yunzhe Li, Yue Ding, Bo Chen, Xin Xin, Yule Wang, Yuxiang Shi, Ruiming Tang, and Dong Wang. 2021. Extracting Attentive Social Temporal Excitation for Sequential Recommendation. In *Proc. CIKM*. 998–1007.
- [34] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proc. WWW*. 689–698.
- [35] Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Noninvasive Self-attention for Side Information Fusion in Sequential Recommendation. In *Proc. AAAI*. 4249–4256.
- [36] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proc. KDD*. 1831–1839.
- [37] Zhiwei Liu, Yongjun Chen, Jia Li, Philip S Yu, Julian McAuley, and Caiming Xiong. 2021. Contrastive self-supervised sequential recommendation with robust augmentation. *arXiv preprint arXiv:2108.06479* (2021).
- [38] Zhiwei Liu, Ziwei Fan, Yu Wang, and Philip S. Yu. 2021. Augmenting Sequential Recommendation with Pseudo-Prior Items via Reversely Pre-training Transformer. In *Proc. SIGIR*. 1608–1612.
- [39] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proc. KDD*. 825–833.
- [40] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proc. KDD*. 483–491.
- [41] Pranava Madhyastha and Rishabh Jain. 2019. On model stability as a function of random seed. *arXiv preprint arXiv:1909.10447* (2019).
- [42] Umaporn Padungkiatwattana, Thitiya Sae-Diae, Saranya Maneeroj, and Atsuhiko Takasu. 2022. ARERec: Attention Local Interaction Model for Sequential Recommendation. *IEEE Access*.
- [43] Apurva Pathak, Kshitiz Gupta, and Julian McAuley. 2017. Generating and personalizing bundle recommendations on steam. In *Proc. SIGIR*. 1073–1076.
- [44] Aleksandr Petrov and Craig Macdonald. 2022. Effective and Efficient Training for Sequential Recommendation using Recency Sampling. *Proc. RecSys* (2022).
- [45] Aleksandr Petrov and Yuriy Makarov. 2021. Attention-based neural re-ranking approach for next city in trip recommendations. In *Proc. WSDM WebTour*. 41–45.
- [46] Michael Potter, Hamlin Liu, Yash Lala, Christian Loanzon, and Yizhou Sun. 2022. GRU4RecBE: A Hybrid Session-Based Movie Recommendation System (Student Abstract). (2022).
- [47] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proc. WSDM*. 813–823.
- [48] Jérémie Rappaz, Julian McAuley, and Karl Aberer. 2021. Recommendation on Live-Streaming Platforms: Dynamic Availability and Repeat Consumption. In *Proc. RecSys*. 390–399.
- [49] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proc. CUAI*. 452–461.
- [50] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proc. WWW*. 811–820.

- [51] Zihan Song, Jiahao Yuan, Xiaoling Wang, and Wendi Ji. 2021. Capturing Multi-granularity Interests with Capsule Attentive Network for Sequential Recommendation. In *Proc. WISE*. 147–161.
- [52] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proc. CIKM*. 1441–1450.
- [53] Zhongchuan Sun, Bin Wu, Youwei Wang, and Yangdong Ye. 2022. Sequential graph collaborative filtering. *Information Sciences* 592 (2022), 244–260.
- [54] Jiayi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proc. WSDM*. 565–573.
- [55] Xiaohai Tong, Pengfei Wang, Chenliang Li, Long Xia, and Shaozhang Niu. 2021. Pattern-enhanced Contrastive Policy Learning Network for Sequential Recommendation. In *IJCAI*.
- [56] Quyen Tran, Lam Tran, Linh Chu Hai, Ngo Van Linh, and Khoat Than. 2022. From Implicit to Explicit feedback: A deep neural network for modeling sequential behaviours and long-short term preferences of online users. *Neurocomputing* (2022).
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*. 5998–6008.
- [58] Chenyang Wang, Weizhi Ma, and Chong Chen. 2022. Sequential Recommendation with Multiple Contrast Signals. *ACM Transactions on Information Systems (TOIS)* (2022).
- [59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
- [60] Qitian Wu, Chenxiao Yang, Shuodian Yu, Xiaofeng Gao, and Guihai Chen. 2021. Seq2Bubbles: Region-Based Embedding Learning for User Behaviors in Sequential Recommenders. In *Proc. CIKM*. 2160–2169.
- [61] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangan He. 2019. A simple convolutional generative network for next item recommendation. In *Proc. WSDM*. 582–590.
- [62] Zhenrui Yue, Zhankui He, Huimin Zeng, and Julian McAuley. 2021. Black-Box Attacks on Sequential Recommenders via Data-Free Model Extraction. In *Proc. RecSys*.
- [63] Zheni Zeng, Chaojun Xiao, Yuan Yao, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2021. Knowledge transfer via pre-training for recommendation: A review and prospect. *Frontiers in big Data* (2021), 4.
- [64] Lingxiao Zhang, Jiangpeng Yan, Yujiu Yang, and Li Xiu. 2020. Match4rec: A novel recommendation algorithm based on bidirectional encoder representation with the matching task. In *Proc. ICONIP*. Springer, 491–503.
- [65] Yixin Zhang, Lizhen Cui, Wei He, Xudong Lu, and Shipeng Wang. 2021. Behavioral data assists decisions: exploring the mental representation of digital-self. *International Journal of Crowd Science* (2021).
- [66] Pengyu Zhao, Tianxiao Shui, Yuanxing Zhang, Kecheng Xiao, and Kaigui Bian. 2021. Adversarial oracular seq2seq learning for sequential recommendation. In *Proc. ICJAI*.
- [67] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proc. CIKM*. 4653–4664.
- [68] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proc. KDD*. 3985–3995.
- [69] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proc. CIKM*. 1893–1902.